

Computational Approaches to Subjective Interpretation of Multimedia Messages

Vom Fachbereich Sozialwissenschaften
der Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
genehmigte

Dissertation

vorgelegt von
Philipp Blandfort

Tag der Disputation: 14. Januar 2020, Kaiserslautern
Dekanin: Prof. Dr. Shanley E. M. Allen
Vorsitzender: Prof. Dr. Thomas Lachmann
Gutachter: 1. Prof. Dr. Shanley E. M. Allen
2. Prof. Dr. Prof. h.c. Andreas Dengel

D 386

Februar 2020

Contents

Abstract	iii
1 Introduction	1
1.1 Overall Goal	2
1.2 Challenges	2
1.3 Research Questions	3
1.3.1 Modeling Interpretation	3
1.3.2 Subjective Image Interpretation	3
1.3.3 Gang Violence Prevention	4
1.3.4 Comparing the Two Application Studies	4
1.4 Contributions	4
1.5 Structure	6
2 Theoretical Background	7
2.1 What is Interpretation?	7
2.1.1 Interpretation in this Thesis	8
2.2 Interpretation and Meaning	8
2.2.1 Meaning of Language	8
2.2.2 Meaning of Visual Contents	9
2.2.3 Meaning in Terms of Implications	9
2.2.4 Implications about Co-occurring Contents	10
2.2.5 Purpose-Specific Implications	10
2.2.6 Meaning in this Dissertation	11
2.3 Factors Influencing Interpretation	12
2.3.1 Priming Effects	12
2.3.2 Cognitive Biases	13
2.3.3 Social Context	13
2.3.4 Factors that are Internal to the Interpreter	13
2.4 Subjective Interpretation	14
3 Technical Preliminaries	17
3.1 Basic Terminology	17
3.2 Supervised, Unsupervised and Semi-supervised Learning	18
3.3 The Supervised Machine Learning Pipeline	19
3.3.1 Building Datasets	20
3.3.2 Common Supervised Machine Learning Models	21
3.3.3 Training and Evaluating Supervised Machine Learning Models	22
3.4 Adopting Supervised Learning	23

4	Related Work	25
4.1	Work on Predicting Subjective Interpretation	25
4.1.1	Information Retrieval	25
4.1.2	Data Mining	26
4.1.3	Computational Communication Research	27
4.2	Modeling Interpretation	27
4.3	Subjective Interpretation of Images	28
4.3.1	Detecting Subjective Visual Concepts	28
4.3.2	Subjective Image Captioning	29
4.4	Gang Violence Prevention	29
5	Own Work	31
5.1	Modeling Interpretation	31
5.2	Subjective Image Interpretation	32
5.2.1	Detecting Subjective Visual Concepts	32
5.2.2	Subjective Image Captioning	32
5.3	Gang Violence Prevention	33
6	Publications	37
6.1	Interpretation Analysis	39
6.2	Fusion Strategies for Learning User Embeddings	95
6.3	The Focus-Aspect-Value Model	103
6.4	Image Captioning in the Wild	113
6.5	Concept and Syntax Transition Networks	123
6.6	Concept and Syntax Transition Networks II	127
6.7	Multimodal Social Media Analysis	133
6.8	Visual and Textual Analysis of Social Media	145
6.9	Evaluating Annotation Disagreement	171
7	Conclusion	181
7.1	Modeling Interpretation	181
7.1.1	Summary	181
7.1.2	Future Work	181
7.2	Subjective Image Interpretation	182
7.2.1	Summary	182
7.2.2	Future Work	182
7.3	Gang Violence Prevention	182
7.3.1	Summary	182
7.3.2	Future Work	183
7.4	Main Research Question	183
7.4.1	Domain-specific Findings	183
7.4.2	Ethics	184
7.4.3	Overcoming Challenges	184
7.5	Overall Lessons	185
	Acknowledgments	187
	About the Author	188
	Bibliography	189

Abstract

Nowadays a large part of communication is taking place on social media platforms such as Twitter, Facebook, Instagram, or YouTube, where messages often include multimedia contents (e.g., images, GIFs or videos). Since such messages are in digital form, computers can in principle process them in order to make our lives more convenient and help us overcome arising issues. However, these goals require the ability to capture what these messages mean *to us*, i.e., how we interpret them from our own subjective points of view. Thus, the topic of this thesis is to advance a machine’s ability to interpret social media contents in a more natural, subjective way. This leads to the main research question:

(Q1) *How can we teach computers to interpret messages with images in a subjective way?*

This ability to handle subjectivity is crucial for applications where user preferences need to be taken into account, such as improving user experience by finding contents of interest or assisting users in formulating social media posts. An example that involves another kind of subjectivity is content moderation. While the huge amount of social media posts has made it virtually impossible for content moderators to check all contents manually, with the ability to automatically detect posts that are perceived as aggressive, problematic posts could be forwarded to someone who can prevent further escalation. At the same time, implementing methods for subjective interpretation forces us to come up with precise modeling approaches, which ultimately contributes to a better understanding of human interpretation. However, subjective interpretation is a challenging topic even for humans, as evident by ongoing debates in the humanities about what it means to interpret (see e.g., recent works in philosophy [1] or hermeneutics [2]). So it is not surprising that automatic subjective interpretation is still far from being solved, despite many relevant efforts from various active research fields including data mining, information retrieval, recommendation systems and content moderation.

In fact, the abundance of related work poses another severe challenge and motivates us to not aim at solving the research question (Q1) in its entirety, but to advance the field by addressing three relevant sub-questions. The first sub-question is concerned with the big picture and deals with general basics and modeling:

(Q1.1) *How to model human interpretation for machine learning?*

In the humanities there is much discussion on how we interpret (e.g., [1, 2, 3, 4]), but no general models are proposed that could directly be used for computation. In computer science, on the other hand, even though “interpretability” is currently a very active research topic, surprisingly it is rarely even made explicit what interpretation means [5]. A notable exception is [6], where interpretation is defined as “mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of”. However, this definition is exclusively meant for computer interpretation, and does not comprise human interpretation. Overall we have identified an

evident lack of transfer between computer science and the humanities. Inspired by recent philosophical treatments on points of view [1], we modify the definition of [6] such that it becomes applicable to both human and machine interpretation and can be more easily adopted in the case where several ways of interpretation are involved. These properties enable many possibilities for computational analysis of single or multiple ways of interpretation within the same theoretic framework, which we collect in a comprehensive survey. Particularly interesting are machine learning approaches where a single neural network learns multiple ways of interpretation, for example by training a neural network to predict user-specific movie ratings from movie features and user ID. This is a promising direction, as neural networks are powerful for prediction, and analyzing what they have learned can help to better understand the different ways of interpretation. However, we point out several open issues related to the reliability of such an analysis. In particular, with the example of movie ratings, we show that the way of combining information for prediction can affect both prediction performance and what the network learns about the various ways of interpretation (corresponding to users).

So far we discussed challenges and opportunities for dealing with human interpretation on a general level. However, subjectivity-related problems can come in diverse shapes and some application-specific details only become visible when focusing on particular problems. Therefore, the rest of the dissertation is concerned with two selected application domains: Subjective visual interpretation and gang violence prevention.

The first application study deals with subjectivity that comes from personal attitudes and aims at answering:

(Q1.2) *How can we predict subjective image interpretation one would expect from the general public on photo-sharing platforms such as Flickr?*

This question is relevant for several use-cases. For example, predicting subjective image interpretation can be used for automatic image tagging or to help users find images according to subjective criteria. Such use-cases require an output that is more informative than a single scalar as common in sentiment detection. Hence, most relevant here is the line of work on detecting subjective adjective-noun combinations (such as “cute puppy” or “scary dog”) from images [7, 8, 9], based on the Visual Sentiment Ontology [7]. We continue this line of research by introducing the Focus-Aspect-Value Model for subjective interpretation of images, which can be seen as a more structured version of adjective-noun-based detection, taking into account semantic relations between adjectives and making it possible to do 0-shot learning. Yet another way to interpret images is to convert them into a phrase or short sentence, that is to caption them. This form of interpretation becomes relevant if, for example, one would like to develop an assistant system that suggests titles for images the user wants to share online. In a crowdsourcing study we analyze image captions on Flickr, revealing that users generally preferred captions with sentiment, which suggests that some subjectivity should be incorporated in image captioning approaches if the goal is to obtain appealing captions. As a first step into this direction we develop a pipeline approach to subjective image captioning, where we first extract adjective-noun pairs as mid-level concepts and then feed them to a language module which turns the detected concepts into a caption such as “mucky and tired baby”.

Our second study is about gang violence prevention and considers the following research question:

(Q1.3) *How can we automatically detect tweets of gang members which could potentially lead to violence?*

Alarming high numbers of gang-related crimes have been reported in several major US cities. For addressing this issue, we aim at assisting local intervention workers by informing them about problematic tweets, so that they can prevent escalation by approaching involved individuals before any aggressive offline action is taken. Since we do not have ground truth information for violent acts linked to Twitter users, this approach crucially involves identifying tweets perceived as *aggressive* by other people in the gang community. This makes apparent the relevance of a more complex kind of subjectivity, which is more indirect and community-specific than in the other case study. A first step into this direction was done by Blevins et al. [10], who built text-based methods for detecting aggressive and loss-related tweets from gang-associated youth in Chicago. However, being entirely text-based, their work ignores images that are often included in tweets and can contain important extra information. Therefore, we extend their work by taking images into account as well. To this end, we created a new dataset of tweets with images, annotated by *aggression*, *loss* and *substance use*, as well as mid-level visual concepts such as *handgun*, *hand gesture*, and *marijuana*. Note that obtaining ground truth labels for codes such as *aggression* is especially challenging and mislabeling can potentially have detrimental consequences. This naturally moves the focus towards annotation. In particular, for such scenarios we recommend in-depth collaboration between computer science and social work or social science, a set-up for which we developed the open-source web-based annotation system VATAS. Additionally, we introduce new strategies for analyzing annotator disagreement, which we find to be beneficial for getting a deeper understanding of the community and which confirmed that subjectivity of annotation was important to consider. Finally, we build multimodal detectors for the three codes and show that using image and text modalities for detection leads to a large relative improvement as compared to any detector using just a single modality.

In summary, in this thesis the overall goal is to advance the ability of computers to interpret multimedia posts in subjective ways, so that they can become more valuable assistants. In two distinct application domains, we develop machine learning models for predicting subjective interpretations of images or tweets with images, respectively. Detection of subjective concepts remains a challenging (and subjective) task, thus we incorporate mid-level concepts into most of these models. Doing so adds an explainability component, helps to guide further progress and even proves useful for analyzing annotation disagreements. This pipeline approach can easily be adapted to other domains. In the process of building these detection tools, we also created three different datasets which we share with the research community. Furthermore, we see that some domains such as Chicago gangs require special care due to high vulnerability of involved users. This motivated us to establish and describe an in-depth collaboration between social work researchers and computer scientists. As machine learning is expanding towards more subjective applications and gaining societal impact, we have good reason to believe that similar collaborations between the humanities and computer science will become increasingly necessary to advance the field in an ethical way. Finally, our work on modeling interpretation helps us to find and structure many possibilities for analyzing interpretation with computational methods, but also makes us realize general open issues concerning the reliability of some popular analysis methods. Since understanding surely is important for guided progress, both of these points will be important to consider for further research on the topic.

Chapter 1

Introduction

Over the past few decades, we could observe a drastic change in everyday communication, as more and more communication went online. In 2018, the global average of daily time spent on social media platforms such as such as Twitter, Facebook, Instagram, YouTube etc. was reported to be almost 2.5 hours [11], which shows how prevalent online communication has become. Remarkably, the amount of multimedia contents involved in these digital interactions has grown to a huge scale, causing researchers to talk about “multimedia big data” (see e.g., [12]). Not only the mere quantity of multimedia messages makes them important, but some findings even suggest that posts with multimedia data are more impactful. For example, tweets with images were found to receive 15% more clicks, 89% more favorites and 150% more retweets as compared to text-only tweets [13].

Unfortunately, the rise of online communication came together with several novel issues. For instance, cases of online violence have been reported [14, 15], and researchers have found excessive social media usage to negatively affect psychological well-being [16]. Such issues are especially challenging to address due to the large scale of related online data. In particular, it is typically not feasible for content moderators to manually inspect all messages between users for taking action against online violence.

On the other hand, the shift toward online communication brought about great opportunities: Since online communication happens in digital form, computers can in principle process such user interactions. This can be helpful for improving user experience, learning more about how we communicate, or even helping us to overcome the previously-mentioned issues.

Indeed, already today we find many useful approaches in such directions: Computer models have been built for making communication more efficient by generating reply messages [17], helping us find contents we would likely be interested in [18, 19], or assessing photo aesthetics for providing instant feedback to users [20]. Similarly, systems have been created for monitoring mental health [21] or detecting aggressive posts such that they can be forwarded to someone who can prevent escalation [10]. What many of these approaches have in common is that at their core they are trying to capture what online messages mean *to us*, i.e., how we interpret them from our own subjective points of view. Coming back to the previous examples, if we are able to predict which posts a given user finds interesting, appealing or aggressive, then several of the above-mentioned applications become relatively simple to solve. Even for the generation of messages it can be argued that a major part of the problem is to judge whether some given statement is appropriate as a response.

As we will come to see in the remainder of this dissertation, there have been considerable advances in automatic processing of online messages during the past few years. Still, while working with subjective interpretation of textual contents has a

longer tradition in machine learning, when it comes to subjective interpretation of multimedia messages, results are lacking behind. Furthermore, even though there are many individual applications which deal with subjectivity, the core topic of how subjective interpretation can be predicted is rarely discussed on a general level.

1.1 Overall Goal

The goal of this thesis is to advance a machine’s ability to interpret social media contents in a more natural, subjective way. This leads to the *main research question*:

(Q1) *How can we teach computers to interpret messages with images in a subjective way?*

In Chapter 2 we will describe at length what exactly is meant by *subjective interpretation*. Until then, the following working definitions shall be sufficient:

- By *interpretation* we mean to assign meaning, which we assume to take the form of categorical labels (e.g., classes such as “authored by XY”, “contains a dog”, “perceived as aggressive”) or vectors (e.g., concept likelihoods), while
- interpretation is called *subjective* if the assigned meaning depends on personal attitudes.¹

For example, interpreting something as “interesting” or “not interesting” is subjective because it depends on personal attitudes. In many cases, deciding whether something is small vs large, light vs dark, etc. is subjective as well. A non-subjective example would be to decide whether an image contains a dog, a car or a horse (unless the image is ambiguous).

1.2 Challenges

Even most humans are well aware of everyday difficulties caused by misunderstanding each other, an issue which essentially is about predicting how someone else would interpret what is being said or done. Thus, intuitively, we already find good reasons for thinking that automatic subjective interpretation is likely going to be a difficult undertaking.

On the computational side, this difficulty is reflected in challenges in obtaining suitable data for building and evaluating machine learning models. In particular, subjectivity requires extra efforts when collecting annotations, and – as compared to purely objective cases – it is less clear what the desired output of a trained model should be.

But the challenges go much deeper than that: If we turn toward the humanities, we find that it is already hard to define many of the central concepts that together make up the topic of this thesis. This is especially true for interpretation: Philosophical efforts to understand interpretation date back more than 2,000 years (see [1]) and we still find no generally accepted description of what exactly interpretation is. In particular, interpretation is closely related to yet another conceptually difficult concept, and that is the notion of *meaning*. Due to these conceptual issues, in Chapter 2 we will explain in detail how we interpret these central concepts of this dissertation.

Moreover, there is a huge amount of relevant literature on the overall topic of subjective interpretation, and many recent efforts in the field of Artificial Intelligence

¹Another type of subjectivity that will be discussed later is a dependency on the relation between the world and the subject(s) that is interpreting. However, this other type is less relevant for this dissertation.

deal with related applications. This abundance of relevant literature and existing efforts poses another severe challenge and makes it unfeasible to solve the research question (Q1) in its entirety.

1.3 Research Questions

In order to keep the scope of this dissertation limited and still take the big picture into account, we aim at advancing the field by addressing three sub-questions. These sub-questions correspond to the topics *modeling interpretation*, *subjective image interpretation* and *gang violence prevention*.

1.3.1 Modeling Interpretation

Interpretation is the most central concept of this dissertation. As we have argued above, interpretation is also a complex concept and can be very hard to grasp. Thus, in order to make any solid advances in predicting subjective interpretation, it is necessary to take sufficient time for looking closer into how interpretation can be modeled for computational approaches. This motivates the research question:

(Q1.1) *How to model human interpretation for machine learning?*

Basically, we propose a supervised learning formulation for interpretation, where interpretation is described as a mapping from input to meaning. Our investigations into these basics allows us to develop a solid mathematical understanding of interpretation. This mathematical clarity leads us to many possibilities for *interpretation analysis*, which we find to be helpful for guiding overall progress and thus we investigate this option in detail.

Note that in our work on modeling interpretation, we are concerned with the big picture and treat interpretation at a very general level. However, subjectivity-related problems can come in diverse shapes and some application-specific details only become visible when focusing on particular problems. For this reason, we also handle two selected application studies (*subjective image interpretation* and *gang violence prevention*), which we introduce next.

1.3.2 Subjective Image Interpretation

Our first application study aims at answering the research question:

(Q1.2) *How can we predict subjective image interpretation one would expect from the general public on photo-sharing platforms such as Flickr?*

More specifically, for this case study we care about subjectivity that often becomes visible through adjectives and comes from typical personal standards, attitudes or preferences. For example, a picture of a dog might evoke feelings of affection in some people and fear in others. Such feelings can become visible in user-generated contents such as image titles (e.g., “our favorite family member”), tags (e.g., “cute dog” or “scary dog”) or comments (e.g., “how adorable!”), where adjectives such as “favorite”, “cute”, or “scary” contain most of the subjective part of interpretation.

All of these types of information (i.e., titles, tags, comments) can in principle be used for building computer models, where the individual types of information lead to different use-cases. For example, predicting subjective tags or other *concepts* for images can be used for automatic image tagging or to help users find images according to subjective criteria. Interpreting images by converting them into a phrase or short sentence – that is to *caption* them – leads to other use-cases. For example, this form

of interpretation becomes relevant if one wishes to develop an assistant system that suggests titles for images the user wants to share online.

In this dissertation, we handle both of these cases, i.e., we look into approaches for predicting subjective *concepts* and approaches for generating *phrases* with a subjective component.

1.3.3 Gang Violence Prevention

Our second application study is about gang violence prevention and considers the following research question:

(Q1.3) *How can we automatically detect tweets of gang members which could potentially lead to violence?*

With this question, we address a severe societal issue the US are facing: Gun violence has reached problematic levels in many major US cities. One of the worst cases is observable in the city of Chicago, where a 58% rise in gun homicides was reported in 2016, involving over 4,000 shooting victims [22]. Victims and perpetrators of gun violence were found to often have gang associations [22], which motivates increased efforts to reduce violent conflicts of gang associates.

Observations that gang members post publicly on social media [15] and that online aggression can lead to offline violence [23, 24] suggest that social media posts of gang associates might be useful for understanding and ultimately preventing violent conflicts from escalating. In fact, this explains why we ask the specific research question of this application study. In our work, we focus on detecting the psychosocial codes *aggression*, *loss* and *substance use*. Since aggression and loss crucially involve evoking and expressing feelings, the relation to subjectivity is evident for two of the three codes.²

1.3.4 Comparing the Two Application Studies

Both application studies mainly aim at detecting subjective concepts from online posts with images. However, there are several important differences between the tasks, which are summarized in Table 1.1: For the two case studies, very different ways of interpretation are important. In particular, this difference is reflected in different output formats, types of subjectivity, and the community/domain that is dealt with. We shall point out that the kind of subjectivity relevant for gang violence prevention is considerably more complex than for our subjective image interpretation study: Judging whether a tweet displays *aggression* means to estimate the effect of tweets on potential recipients within a specific community. Hence, for gang violence prevention we are dealing with a very domain-specific and challenging kind of interpretation. This naturally moves the focus of the study to annotation and domain understanding, which are far less critical for subjective image interpretation. The disparity between the two studies will turn out to be useful in order to see different points that are relevant for understanding the broader topic of subjective interpretation.

1.4 Contributions

The main contributions of this thesis can be summarized as follows:

²Whether a tweet is related to *substance use* can also be a subjective question. For example, there is not necessarily a definite objective answer to the question whether a Styrofoam cup in an image contains the drug *lean* or not.

	Subjective Image Interpretation		Gang Violence Prevention
	Concept-based	Phrase-based	
Modalities	image only		image + text
Platform	various websites	mostly Flickr	Twitter
Community	general public		Chicago gangs
Input	images		tweets with images
Mid-level Concepts	–	adjective-noun combinations	9 domain-specific concepts
Output	structured visual interpretation	image captions	psychosocial codes
Labels	weak, user-generated		manually annotated
Subjectivity	attitude-based, direct		attitude-based, indirect
Sources of Subjectivity	common attitudes, preferences		emotions, ambiguity
Use-cases	data mining, information retrieval	entertainment, convenience	gang violence prevention (social work)
Focus	modeling, prediction	explainability, prediction	annotation, domain understanding, ethics

Table 1.1: Comparison of the two application studies. Due to their differences, our application studies shed light on different facets of building models for predicting subjective interpretation.

1. *Modeling*: We propose a general way of modeling human interpretation that can directly be used for machine learning and proved valuable for finding and structuring methods for analyzing interpretation. Further, our definition of interpretation reveals a relation to the mathematical field of Functional Data Analysis, which brings about new opportunities for dealing with interpretation-related data.
2. *Analysis*: Based on our proposed modeling of interpretation, we formally describe the problem of *computational interpretation analysis, survey* and *structure* relevant approaches. In doing so, we describe a broader picture than related overview papers [6, 5, 25, 26] and point out open issues regarding model-based analysis. In particular, we describe strategies for comparing multiple ways of interpretation which reveals new possibilities for analysis.
3. *Detection*: We develop various machine learning models for predicting subjective interpretation of images and tweets with images, which were able to improve over previous state-of-the-art (tensor fusion for subjective image interpretation, multimodal detection model for gang violence prevention) or add an explainability component (CAST for subjective image captioning).
4. *Annotation*: For many machine learning approaches it is necessary to have annotated data for training and evaluation. We describe an interdisciplinary collaboration for high quality annotation, propose strategies for analyzing annotation disagreements, and release an open-source annotation system. These

contributions are made in the context of gang violence prevention but can be generalized to other domains.

5. *Datasets*: We build and release three datasets (image captioning annotations from crowdworkers, *aspects-DB* for subjective image interpretation, and annotated tweets of gang-involved youth) to the research community. These datasets can be used to reproduce our findings, do further analyses and run experiments with new models.

1.5 Structure

The contents of this dissertation are structured as follows: Chapter 2 lays out the theoretic grounding of the presented work by defining subjective interpretation and discussing related concepts. Chapter 3 then introduces the necessary machine learning basics, primarily for readers without a strong technical background. Afterwards, in Chapter 4 we have a look at existing computational approaches to subjective interpretation. In Chapter 5, we then summarize our own approaches to interpretation modeling and the two application studies. Chapter 6 includes all publications that constitute the main part of this dissertation. Finally, we conclude in Chapter 7 with a summary of the main findings, a discussion of open issues and prospects of promising future work.

Chapter 2

Theoretical Background

For this dissertation, it is crucial to understand what *subjective interpretation* is supposed to mean. In this chapter, we take a step by step approach for introducing our notion of subjective interpretation, starting by defining interpretation.

2.1 What is Interpretation?

When we look up the definition of the word “interpretation” in the dictionary Merriam-Webster, we find that it is defined as “act or result of interpreting” [27]. The main concept here is the verb “to interpret”, for which the same dictionary lists two senses [28]:

1. “to explain or tell the meaning of : present in understandable terms”, as in “interpret dreams” or “needed help interpreting the results”
2. **“to conceive in the light of individual belief, judgment, or circumstance : construe”** as in “interpret a contract”

For this dissertation, the focus lies on the second sense, where interpretation describes a more or less automatic process we deploy for making sense of the world as we perceive it. Note that this second sense is more fundamental than the first, since whenever anyone (or anything) is to explain the meaning of anything, he/she/it first needs to conceive it.

Furthermore, meaning is crucial for this second sense as well and will thus be central for our investigations: If someone construes A as B, this can be understood as the person assigning *meaning* B to A. This relevance of meaning is supported by definitions of *interpreting* in other dictionaries, such as “to understand (an action, mood, or way of behaving) as having a particular meaning” [29] or “to decide what the intended meaning of something is” [30].

Remark: Interpretation and Philosophy

The definition we highlighted above already points to the relevance of situational and personal context for interpretation. This connection is of central importance in the field of *hermeneutics*, which originates from studies of biblical texts, but is nowadays concerned with how we interpret information in general. In hermeneutics it is commonly argued that in order to make sense of things we need to relate them to our own life situation, which makes all interpretation something inherently personal (see e.g., [2]).

This statement, of course, has strong philosophical implications. In particular, it begs questions such as “Is there any objective reality that is independent of interpretation?”, “Do absolute truths exist?” and “Can we judge whether one or another way of interpretation is better?” Questions of that kind have been asked in philosophy for a long time. In fact, the role of interpretation in philosophy goes back at least to around 500 BC, when the philosopher Heraclitus claimed that there was no objective reality, only our subjective “appearances”. The philosophical positions skepticism, relativism and perspectivism all crucially involve the notion of a point of view from which we interpret [1], and ponder on implications of this interaction. Relativism, perspectivism and similar positions still receive much attention in philosophy [1, 31], but are also relevant to present-day debates in politics about topics such as tolerance and cultural diversity.

Such treatments typically do not include any definitions of interpretation that can be used for practical applications. Still, philosophical discussions in [1] about what subjective points of view are will become relevant later on when defining subjectivity (Section 2.4). For now, what we take from studies in hermeneutics and philosophy is the insight that there needs to be someone or something that is interpreting. In other words, there is no interpretation without an interpreting entity.

2.1.1 Interpretation in this Thesis

We have seen that interpretation is one of the most fundamental topics. Based on the common understanding of the term, we can see interpretation as a form of transformation, or conversion of information to something we can refer to as *meaning*.

2.2 Interpretation and Meaning

So, essentially, predicting interpretation means to predict the meaning someone (or something) assigns to information, which brings up the question “What is *meaning*?”

Since we are ultimately interested in interpretation of messages consisting of text and/or images, we now first have a look at how meaning of language and visual information is commonly treated in scientific studies. From there we move to the notion of meaning that will be adopted for this dissertation.

2.2.1 Meaning of Language

The meaning of language has been extensively studied in linguistics (see e.g., the book “Meaning in Language” by Cruse [32]). There are mainly two linguistic fields studying meaning: *Semantics* and *pragmatics*.

Semantics includes several branches, each dealing with a different kind of meaning. To mention some of them: In *lexical semantics*, the intuitive meaning of individual content words (i.e., nouns, verbs, adverbs, adjectives) is analyzed. In particular, such analyses result in descriptions of word senses and relations between words (e.g., synonyms, hypernyms, entailment). On the computational side, we can find related tasks such as word sense disambiguation (i.e., given a word in a context decide which particular word sense is most suitable) and resources such as the popular WordNet [33], which groups together words into sets of synonyms and contains information about word senses as well as word relations.

Grammatical semantics studies relations between meaning and syntax. This includes the meaning of grammatical morphemes like *-ed* in *walked*, or variations of meaning depending on syntactic role (e.g., *white* in “She was dressed in white” vs “She was wearing a white dress”).

Formal semantics uses logical formalism to model the meaning of language (typically propositions or sentences). Basically, we can understand interpretation in terms of formal semantics as a conversion of text to such logical forms.

The line between semantics and *pragmatics* can be quite fuzzy. One way to differentiate between the two is to draw the line based on different uses of the verb *mean* (see [34]): While semantics aims at answering “What does X mean?”, pragmatics is concerned with the question “What did you mean by X?” Thus, *pragmatics* deals with speaker intent, implicatures and other aspects of meaning that depend on context, such as ambiguity that requires situational context to resolve (e.g., “I went to the bank”).

Results in pragmatics can be in the form of rules, such as the famous cooperative principle by Paul Grice, which states: “Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.”¹ [35]

2.2.2 Meaning of Visual Contents

When it comes to visual contents such as images, it is rarely discussed how meaning should properly be defined. Even intuitively, it is less clear what the meaning of an image could be as compared to the meaning of a sentence or word.

In computer vision, semantics of an image are typically described in terms of objective contents that are visible in various areas of the image (see e.g., [36]). In particular, *semantic segmentation* refers to the task of assigning image pixels to concepts (such as *car*, *house*, *street*, *person*) in order to partition the image into coherent parts which capture its main contents. A more complete representation of an image’s meaning is aimed at in the work of Frank Keller [37], who proposes a structured representation of images that also considers relations between visual elements, similar to dependency trees we find in semantics.

Still, as *a picture paints a thousand words*, it might not come as a surprise that the meaning of visual contents can be more complex than what is accounted for in such objective approaches. In case of paintings this becomes very clear, but even photos commonly posted on social media (e.g., of friends, pets or events) evoke reactions on the viewer’s side (including emotions and associations) or have other implications which we can see as part of the meaning. Thus, modeling image semantics in terms of objective contents and their relations can be a useful stepping stone, but especially for images that are part of a message, we argue that implications are crucial to consider. For example, if a friend sends you a close-up of a funny dog face after you mentioned that you are feeling stressed out, segmenting the whole image as “dog” would miss the point entirely.

2.2.3 Meaning in Terms of Implications

In Zimmermann’s introduction to hermeneutics [2] we find the following quote from the Scottish philosopher John Macmurray: “If we did not know what water is by drinking it and boiling it in our kettles, the scientific statement that water is H_2O would be merely a meaningless noise.” What this statement is meant to point out is that meaning is ultimately built on experiential relations, or in other words, on implications for one’s personal life. In the example of the chemical formula for water, in this sense we could say that the meaning of “water is H_2O ” mainly comes from the possibilities it implies, such as the implication that water can be made out of oxygen and hydrogen etc.

¹This statement is often broken down into four maxims of conversation, the so-called Gricean maxims.

Of course, the formulation “implications for one’s personal life” is too general for computational purposes, but we do find studies on the following more specific types of implications (which are applicable to multimedia messages):

- *Co-Occurring Contents*: Implications about which other contents (typically words) can be expected to occur in the same context.
- *Purpose-Specific Implications*: One might want to consider messages as having the same meaning if they lead to the same outcome, given the context or purpose of interpretation.

We will now look a bit closer into each of these types of implications.

2.2.4 Implications about Co-occurring Contents

In linguistics, a popular claim is the *distributional hypothesis* which says that “words which are similar in meaning occur in similar contexts” [38]. Some more radical versions of it have been put forward, e.g., by Baroni and Boleda who stated that “The meaning of a word is the set of contexts in which it occurs in texts” [39]. Surely not every scholar would agree with this claim, but occurrence contexts certainly have a strong relation to meaning. In particular, this becomes clear when inferring the meaning of novel words/symbols. For example, what does the word “wampimuk” mean in these sentences:

1. Ugh, I think I had too much *wampimuk* last night!
2. The other day when I was walking through the woods, I saw a *wampimuk* sitting on a tree.
3. Donald was running late for his appointment at the *wampimuk*.

Based on this idea, in the field of *distributional semantics* (which belongs to natural language processing), meanings of words are estimated based on the contexts in which they occur. Originally, this was done by representing the meaning of a word as a vector which counts how frequently other words occur within the same sentence (given any large corpus). More recently, prediction-based approaches such as the popular word2vec [40] have become dominant. However, the basic idea still remains to describe the meaning of a word in terms of co-occurring contents. There are several variations regarding which contents are taken as co-occurring, such as only considering words which co-occur within some fixed distance (specified as number of words). These variations affect what exactly is inferred as meaning. For instance, only considering words that are direct neighbors leads to a more syntactic form of meaning.

In principle, instead of co-occurring contents, there are many other ways for understanding and defining *context*, which could also be used to capture the meaning according to the distributional hypothesis. In a broad sense, we could consider any other implications as part of the context, which brings us to the next point.

2.2.5 Purpose-Specific Implications

We can understand purpose-specific implications as a special case of context-based approaches to meaning. The idea is to first determine an application or purpose for interpreting (i.e., assigning meaning), and then define meaning in terms of implications for this particular purpose.

For example, if we want to send an image to a friend in order to inform him or her about what kind of car we bought, then all images where the car is clearly visible could be said to have the same meaning. Of course, one might argue that other factors such as the emotional reaction of the friend need to be taken into account as part

of the images’ meaning as well, so that the described equivalence might only be an approximation. Still, more generally, if we think about information as messages in a communication context, it makes intuitive sense to claim that any two messages which have exactly the same effect also must have the same meaning. This idea leads to a notion of meaning that is relatively straightforward to use for computational approaches.

For example, if we want to recommend images to users, the meaning of an image can be taken as the user’s sentiment toward the image, which is observable in form of user interactions (e.g., whether the user views the image or likes it). In this specific context, we might say that what an image means can be sufficiently captured by a binary value indicating like vs not like.

At first glance, such representations of meaning might seem like an oversimplification of the matter. However, meaning in all its semantic and pragmatic subtleties is not even fully comprehensible to humans which suggests that in any case we need to rely on some kind of approximation for computational purposes. Furthermore, when dealing with interpretation for a specific purpose, what exactly would be the benefit of considering aspects of meaning that do not have any direct implications for the purpose at hand?

We shall add that in our application studies we do face situations where the purpose is very abstract or general (e.g., predicting subjective image interpretation) or one has only limited information available (e.g., analyzing online violence on twitter but no emotional reactions to tweets are accessible). Hence, in practice, one might need to come up with *discriminative* concepts that can be observed and relate to the purpose. For example, for subjective interpretation of images we could say that the top adjective someone associates with an image certainly captures some part of the person’s subjective interpretation and is discriminative (since it is very unlikely that the person would use the same adjective for all images). In the example of online violence, it would be possible to introduce tweet labels that are hypothesized to play a role in online violence, such as labels for whether a tweet is perceived as “aggression”, expresses “loss” etc.

To give a few more examples of other work with relations to purpose-specific meaning, we come back to linguistics. Several linguistic branches study variation in language, which revealed several dependencies between language choices and aspects of the author. For instance, *stylistics* studies style in language and how style depends on factors such as genre, author or historical period [41]. Results from stylistics have applications in fields such as *forensic linguistics*, where the goal is to determine the author of texts in legal cases. For this purpose, we could say that the meaning of a text is mainly given by the identity of its author. A related field is *sociolinguistics*, which analyzes how linguistic choices relate to group membership and other social factors. Here, what language means is described by the social context it is used in.

2.2.6 Meaning in this Dissertation

Meaning in the sense of *intuitive understanding* is very complex and not fully graspable. Automatically predicting this kind of meaning amounts to letting the computer perceive the world in a human way, including all the non-conscious nuances and impressions. We will not attempt to do that for two reasons: First, doing so is too ambitious for a dissertation. Realistically, we could only expect to get a very coarse approximation to this type of meaning by computational means. Second, for most practical purposes of developing computer assistant systems with ability to handle subjectivity, it is simply not necessary to model the whole of human interpretation. As it turns out, we may define interpretation in terms of a more restricted form of purpose-specific meaning, which is still powerful enough for any practical applications and cuts off all the unnecessary overhead.

Thus, in this dissertation we use a notion of meaning that depends on the purpose of interpretation. In short, we define meaning in terms of implications for a given purpose, which is typically described by discriminative concepts such as class labels (see Subsection 2.2.5).

Another benefit of this choice is that our notion of meaning is applicable to any type of information, including text, images and multimodal messages. Furthermore, our definition shows a possibility to relate behavior and interpretation, which makes interpretation observable (and measurable). Indeed, with a purpose-specific notion of meaning it is evident how (re)actions can often serve as a proxy to interpretation. For example, if a user can choose between the three reactions “laughing”, “crying”, “surprise” to a post, the chosen reaction can be seen as the user’s interpretation of the post. In particular, we can now also see how experiments in psychology and cognitive science on reactions to stimuli can be understood as analysis of interpretation – something that is hard to see when adopting a semantic notion of meaning. Take Pavlov’s famous conditioning experiments for example [42], where saliva production could be said to indicate which meaning the dog assigns to stimuli such as a ringing bell.

2.3 Factors Influencing Interpretation

Before we define subjectivity, we shall look more generally into which factors have been identified as having an impact on the process of interpretation. In doing so, we do not aim at an exhaustive overview but mention the most prominent cases.

2.3.1 Priming Effects

Priming is a well-studied psychological phenomenon. The underlying idea is that any contents we process stay active in our subconscious mind for some time, and during this time affect processing of successive contents. This influence can be measured either in terms of increased processing speed of related contents, or as increased likelihood for the generation of similar contents. For example, the word “roof” is read more quickly when presented after “house” than when following “water”. This case falls under *semantic priming*, because the boost in processing speed is due to semantic relatedness of the stimuli.

There are various other forms of priming, including syntactic priming. As an example of *syntactic priming*, participants in a description task would use passive verb forms more often after having heard someone use a passive verb form shortly before [43].

Priming effects have been reported for non-textual stimuli as well. For instance, Dell’Acqua and Grainger analyze semantic priming using pictures as stimuli [44]. To mention a more exotic case of priming, Bargh et al. describe an experiment where age-related words mentioned during an interview resulted in slower movement after the interview [45].

It shall be noted that not all experiments on priming describe effects on interpretation outcomes: Processing speed is a property of the interpretation process and not part of what the stimulus means to the participant. If, however, the analysis focuses on changes of participants’ decisions or sentence generation, we can understand priming as an effect of the first stimulus on the meaning assigned to the succeeding one.

2.3.2 Cognitive Biases

In the field of cognitive science, many influences on human information processing and decision making have been collected under the umbrella term *cognitive biases*. (Consider e.g., [46, 47, 48].) Not all such biases apply to the interpretation of online messages. Some prominent examples that do are:

- *Attentional bias* describes the influence of recurring thoughts on attention. For instance, anxious people are more likely to attend to threatening stimuli [49].
- The *Bandwagon effect* refers to the tendency to adopt views and behaviors that are adopted by many others.
- *Confirmation bias* connotes the seeking and interpreting of information such that it supports existing beliefs, expectations or hypotheses [50].
- *Framing effect* refers to the difference in interpreting information depending on the way it is presented [51].
- The *fundamental attribution error* is the tendency to overestimate the importance of people’s character or intention over environmental factors when interpreting the behavior of others [52].
- *Group attribution error* describes the tendency to assume that decisions of a group reflect the attitudes of individual group members [53].

In particular, cognitive biases can be important to consider when annotating data. For annotation with crowdsourcing, Eickhoff showed that common cognitive bases (ambiguity effect, anchoring, Bandwagon effect, and Decoy effect) can indeed affect annotation quality [54].

2.3.3 Social Context

Especially for information in the form of messages, social context is an important influential factor. There is some overlap with cognitive biases and influences that count as social context, such as authority bias (tendency to put more trust in opinions of authority figures) [55] or social desirability bias (tendency to portray oneself as socially desirable) [56, 57].

More generally, it makes intuitive sense to think that the identity of the author can have a strong impact on the interpretation of the viewer. For instance, if a message comes from your mother your interpretation would likely be very different as compared to your interpretation of the same message coming from some complete stranger. Similarly, situational context plays a role in interpretation. For example, if a colleague told you loudly “I think I should clean up my room” during a business meeting with other people present, social implications of the message (which can count as part of its meaning as we argued above) would be different than in a more private context.

Effects of this kind are mainly addressed in communication theory [58, 59] and psychology [60]. Other studies in similar directions include analyses of messages’ implications in pragmatics, and investigations into variation of language according to social factors [61].

2.3.4 Factors that are Internal to the Interpreter

There are various other influencing factors that are internal to the person who is interpreting, including personal preferences and emotional states.

For preferences, the relation to interpretation is most obvious. In fact, the word “preference” is normally defined in terms of interpreting something as more favorable than something else (e.g., as “act of preferring” [62]).

Regarding emotions, in the field of mental health care we find many accounts for influences of emotions on general cognition. In particular, the term *cognitive distortions* refers to irrational thought patterns, which can largely be understood as biases in interpretation of situations and events, and are characteristic to states of anxiety or depression [63, 64]. Additional evidence for emotional influence on interpretation comes from neurobiology, where Antonio Damasio has investigated emotional influences on decision making (e.g., [65, 66]). It is worth mentioning that emotions can also be seen as part of the interpretation outcome, especially when adopting our purpose-specific notion of meaning. For instance, the emotion a message evokes is crucial to consider for applications in entertainment, content moderation or violence prevention.

2.4 Subjective Interpretation

Moving to subjective interpretation, which of the influences from the previous section do we consider to be subjective? Let us first have a brief look into what subjectivity is and how it can be defined. In recent philosophical discussions by Gutiérrez and Campos [67], we find a definition of subjectivity specific to interpretation which we adopt for this dissertation. Using our terminology, we can summarize the two senses of subjectivity they mention as follows:

1. *Subjective impregnation from attitudes*: Meaning can be subjective in the sense that it is influenced or characterized by personal preferences, attitudes, emotions or experiential qualities. For example, colors or smells are subjective in this sense. Other examples would be interpretation in emotional terms (such as interpreting an image as scary, soothing, happy etc.) or in terms of preferences (such as judging aesthetics of images).
2. *Being relative to a certain position of a subject*: Interpretation is subjective in this sense if the meaning resulting from interpretation can only be understood when taking into account the emplacement or position of the subject that is interpreting. For example, if an image of a dog is interpreted as “image of my dog”, this interpretation only makes sense when considering the relation between the dog and the person who is interpreting. Interpretations that include any other spatial, temporal or personal relations involving the interpreter are subjective in this sense as well.

Importantly, we see that interpretation can be subjective not only due to subjective influences on the interpretation process but the meaning itself can be subjective as well. Meaning can be subjective in both of these two senses, i.e., it can be based on attitudes and/or depend on the position of the interpreting subject(s).

For example, if an image is interpreted in terms of pretty vs ugly, the result of interpretation must be subjective in the first sense. Similarly, rating emotional qualities of contents – such as judging whether a tweet is aggressive or displays loss – typically involves personal attitudes. Interpreting something in terms of young vs old is subjective in the second sense, since age is based on the temporal relation of something to the interpreter at the time of interpretation. (In addition, evaluating age can depend on attitudes.)

Coming back to the influences on interpretation from the previous section: Since the second sense of subjectivity is described as property of the interpretation result, influences on interpretation can only be subjective in the first sense. Thus, internal

factors such as preferences and emotions (Subsection 2.3.4) are clearly subjective, while social context (Subsection 2.3.3), priming effects (Subsection 2.3.1) and cognitive biases (Subsection 2.3.2) are considered as being subjective influences only if they involve attitudes of the person interpreting. For priming effects this is normally not the case. Social context, on the other hand, often relates to attitudes and emotions. For some of the cognitive biases it is debatable whether they involve personal attitudes.²

For the most part of this dissertation, we deal with subjective interpretation where the meaning itself is subjective in the way that it depends on personal attitudes. Still, in some of our work (movie ratings experiment and disagreement analysis in our gang violence study) we additionally consider subjective influences on interpretation.

²For this dissertation it will not be necessary to draw a precise line between subjective and non-subjective influences.

Chapter 3

Technical Preliminaries: Supervised Machine Learning

Most of the work presented in this dissertation belongs to the general topic of *machine learning*, which is a subfield of Artificial Intelligence where statistical models and algorithms are applied to data in order to “learn” (i.e., acquire abilities, such as recognizing if there is a dog on an image or not).

In this chapter, we shall briefly review the necessary basics of machine learning. During this line of action we also introduce the computer science terminology for readers from less technical backgrounds, given that most of our publications (Chapter 6) were written for a computer science audience and assume these basics to be known.

3.1 Basic Terminology

In order to understand the contents and contributions of this dissertation, it is necessary to be familiar with several basic concepts from machine learning. The most important ones are listed in Table 3.1. In the remainder of this chapter, we will see

Term	Definition
<i>corpus</i>	systematic collection of naturally occurring data (e.g., texts, images, social media posts)
<i>model</i>	statistical method that maps <i>input</i> samples to <i>output</i> , where the output of the method depends on internal parameters; for example, linear regression that maps from age and gender to estimate body height
<i>prediction</i>	synonym for output of a model
<i>ground truth</i>	information about which output is considered to be correct for certain input samples; for example, actual measurements of people’s body height for various combinations of age and gender
<i>label</i>	ground truth for a single input sample; for example, measurement of a single person’s body height (with known age and gender)
<i>dataset</i>	systematic collection of data samples (e.g., images); often the samples are paired with corresponding <i>labels</i> or <i>ground truth</i> information (e.g., sentiment of images)

Table 3.1: Basic machine learning terminology.

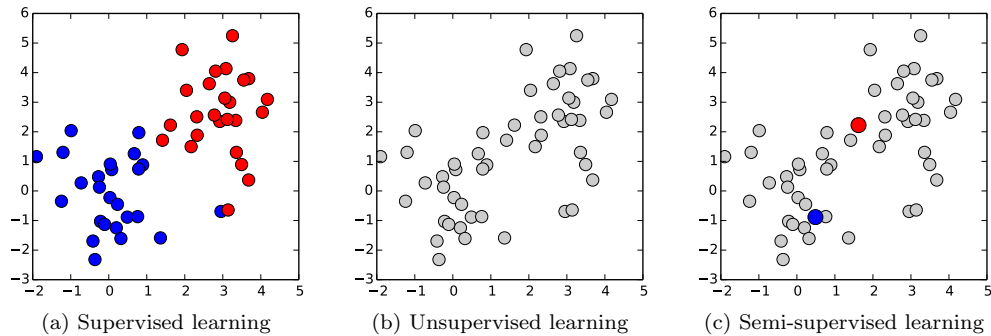


Figure 3.1: Example to illustrate the difference between supervised, unsupervised and semi-supervised learning. A supervised training task would be to classify 2-D points into the colors red and blue, whereas training data we would be given a list of points with known color (a). In case of unsupervised learning there are no classes (b), thus the task could for example be to cluster the given 2-D points into two groups based on their distribution. For semi-supervised learning, the task is the same as for supervised learning, while ground truth information is only available for a small part of the points (c). As an intermediate step, a semi-supervised learning approach might separate the points into two groups and then use the labeled samples to assign these groups to colors. (Figure best viewed in color.)

in more detail how these concepts relate and fit into to the topic of this dissertation.

3.2 Supervised, Unsupervised and Semi-supervised Learning

Machine learning approaches are commonly distinguished based on their degree of “supervision”, which describes how much ground truth information is available for learning. On the extremes we find *supervised learning*, where ground truth information is available for all samples, and *unsupervised learning*, where no ground truth is available at all. The case where only some (typically small) subset of the data has ground truth is referred to as *semi-supervised learning*. This major difference between these paradigms is illustrated in Figure 3.1.¹

Within these three categories we can further distinguish between different machine learning *tasks*:

- Depending on the type of ground truth information, we can distinguish between different *supervised learning* tasks: If we have one or several class labels for each item, the corresponding task to train a model for predicting the class label(s) given an item is called *classification*. For example, a classification task would be to predict whether Shakespeare, Einstein or Trump authored a given text, based on texts coming from these three individuals. In this example, we would say that we classify texts into the *classes* “Shakespeare”, “Einstein” and “Trump”. If the goal is to predict values on a continuous scale, the task is called *regression*. For example, if we are given information about age, gender and height for several people, a regression model could be trained to predict the height of a person

¹There are machine learning approaches which cannot easily be put into any of these three categories, including *reinforcement learning* (e.g., [68]), *active learning* (e.g., [69]), *meta-learning* [70] and *synchronization approaches* such as classless learning [71].

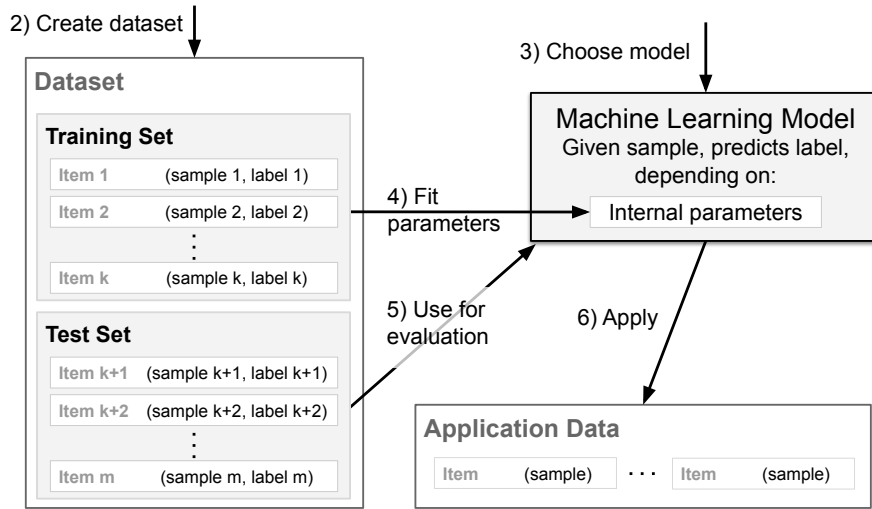


Figure 3.2: Schematic illustration of supervised machine learning with a single model, assuming that a task has already been determined. First, a dataset is chosen or created. Then, a machine learning model is chosen. Next, the model is trained on the training part of the dataset. Afterwards, the test set is used for evaluating the model. Finally, the trained model is applied. Note that this application step is typically not part of research anymore.

based on age and gender. In case the output takes a more complex form such as sentences or images, we talk about *generation* tasks. An example would be to generate titles for images, given a dataset of images together with suitable titles.

- Tasks in *unsupervised learning* generally aim at finding and describing statistical properties of data. The most common task is *clustering*, where the data is automatically partitioned into groups called clusters, such that similar items fall into the same cluster. For example, given a collection of 10,000 images, one might be interested in organizing them into 100 groups such that images in any group are similar to each other. This can be useful in order to get an overview of the image collection, since looking at a few examples from each cluster can already give a good idea of its contents. Other tasks such as *density estimation* are focused more directly on statistical properties of the data.
- *Semi-supervised learning* generally aims at solving the same tasks as in supervised learning and uses unlabeled samples as auxiliary data.

This thesis is mainly concerned with classification and generation, which both belong to supervised learning. In fact, we will see that our chosen approach for modeling interpretation can be understood as supervised learning formulation. As we adopt a supervised learning approach for predicting subjective interpretation, we will now recall the basics of supervised machine learning.

3.3 The Supervised Machine Learning Pipeline

The basic supervised learning pipeline is shown in Figure 3.2. The main steps are as follows:

1. Decide which task you want to solve (e.g., formulate a research question or set a specific application goal) and determine how it can be computationally *modeled*. For example, one might want to work on a common existing task such as sentiment detection, or start with the question of how subjective impressions of images can be captured automatically. The most crucial parts of modeling are to decide on what to use as input (e.g., single images), output (e.g., adjective-noun combinations visible in the image) and find a suitable way of evaluation (i.e., quantifying how well a given model can solve the task).
2. A *dataset* needs to be obtained. In supervised learning, a dataset needs to contain samples and corresponding labels. The labels can be of various types, including categorical (e.g., cat, dog, person), numerical, sequential (e.g., text), while this type is assumed to be the same for all samples of the dataset. We can think of the dataset as a list of examples that describe the desired behavior of the model.
3. Choose one or several machine learning *models*. The models can be any statistical model (e.g., logistic regression, neural networks, support vector machines) that has internal parameters and estimates likelihoods for the possible labels given as input any particular sample.
4. Once a dataset is in place and the model(s) has/have been chosen, a part of the dataset – the so-called *training set* – is used to *train* the machine learning model(s) to predict the correct label given an input sample. In machine learning, training or learning means to optimize the model parameters such that the overall probability of observing the combinations of samples and labels in the training set is maximized.
5. After the models have been optimized, they are *evaluated* using another portion of the dataset, the so-called *test set*. This evaluation is meant to give an idea of how well the models can be expected to perform on unseen data.
6. Finally, the model is *applied* in the scenario it was designed for. This, however, is normally not part of research anymore and not relevant for this particular dissertation.

We will now have a closer look at the parts that are especially relevant to our publications.

3.3.1 Building Datasets

In the process of this dissertation, we built and released several datasets. Especially in our application study on gang violence prevention, annotation plays a central role.

Data Acquisition

For our work, input samples are generally in the form of user-generated contents, so the starting point is either an existing dataset or scraping data from online platforms. Scraping can be done in various ways, including searching for contents that mention or are tagged with specific *keywords*, and obtaining posts from given *authors*. This leads to a collection of data samples. To have a full dataset for supervised learning, another ingredient is required, namely ground truth information.

Ways of Assigning Ground Truth

The most common ways of obtaining ground truth information are manual annotation, automatic annotation and scraping. We will briefly go through all of these options with a simple example for illustration. So let us assume we want to build a dataset with dog vs cat images.

- *Manual Annotation:* In manual annotation, we start with a set of images, where for a given image we do not have any information about whether there is a dog or a cat on the image. We would then ask human annotators to add this information by ground truthing the images and specifying for individual images whether a dog or a cat is shown. These annotators can either be a group of individuals who are registered on so-called crowdsourcing platforms (e.g., Amazon Mechanical Turk), or in-house annotators (e.g., students).
- *Automatic Annotation:* Another possibility is given by automatic annotation, where computer tools such as existing machine learning models are employed to add labels to the data. This is quite common to do for datasets in natural language processing, where parts-of-speech taggers are often used to add parts-of-speech information to texts in the dataset. However, as this approach requires a working system for generating the labels, it is only appropriate as auxiliary information.
- *Using Metadata:* At times, the acquired data contains metadata that can be directly used as labels. For example, for our set of images we might have user tags available and these include the words “cat” and “dog”. On the downside, metadata of user-generated content typically contains *noise*, that is inaccurate or irrelevant information, which often necessitates additional strategies for cleaning up the data. Note that if metadata is used as ground truth, this choice should usually be taken into account already during the data acquisition phase. For example, one would search specifically for images tagged with either “dog” or “cat” when trying to build a dog vs cat classifier.

In the process of this dissertation, we create datasets under the use of metadata, crowdsourcing and in-house annotation.

3.3.2 Common Supervised Machine Learning Models

Even though in several of our papers we mix and compare various machine learning models, in most cases it will be sufficient to treat the upcoming models as black-boxes. Thus, we will not go into the heavy mathematical details but aim at providing a more intuitive idea of the differences between the models.

Some of the common models used in our publications are:

- *Linear/logistic regression* approximate the output as linear combination of the input components. For example, given height and weight of a person, the age of the person could be estimated as a *linear combination*

$$\text{age} = a \cdot \text{height} + b \cdot \text{weight} + c,$$

where a, b, c are real-valued parameters that are estimated based on a given dataset. In case of logistic regression, the logistic function is applied to the linear combination to map the value to the range from 0 to 1. (In the previous example, this would make sense if we wanted to classify into “under 18” and “18 years or older”.) Linear and logistic regression are fairly robust and simple to apply, but by design these models are not capable of learning non-linear

dependencies² which leads to poor performance when applied to more complex data.

- *Support Vector Machines (SVM)* offer more complexity than linear models such as logistic regression. At the same time, SVMs are simpler to train than for example neural networks. On the downside, they are less flexible than neural networks, which makes them more dependent on characteristics of the input (input features) and less suitable for combining various types of information for prediction.
- *Neural networks* are statistical models which are organized into layers. There are various types of layers but most of the heavy lifting in neural networks is done by a sequence of linear transformations followed by an activation function (e.g., logistic function). Hence, roughly speaking, we can think of basic neural networks as combinations of many logistic regression models. Given sufficient network size, even networks with a simple structure were shown to already be complex enough to approximate any arbitrary multivariate function [72]. Recently, neural networks have become very popular as they reached state-of-the-art performances in many machine learning tasks such as image classification [73], machine translation [74] and sentence classification [75]. Advantages of neural networks are that they scale well with dataset size, can achieve good performance with almost any kind of input (e.g., image pixels, frequency spectrum for audio, sequence of words) and are very versatile. However, this flexibility comes with the price of making it very time-consuming to optimize any non-standard network.

3.3.3 Training and Evaluating Supervised Machine Learning Models

Assume we are in a situation where we have a dataset (including samples and associated ground truth), have picked a machine learning model and now wish to train the model with the available data.

Splitting the Dataset

Remember that the goal of supervised machine learning is to obtain a model that is able to predict labels for unseen input samples. For being able to get a fair idea of how well the model can be assumed to generalize to unseen data, it is necessary to split the dataset into a *training set* which is used for optimizing the model, and a *test set* based on which the model is then evaluated.³ If hyper-parameters (such as numbers of layers for neural networks) are optimized or a stopping criterion is necessary for knowing when training was enough, an additional *validation set* might be used.

Training

Optimizing the model is typically done by using a gradient-based approach like stochastic gradient descent. The idea behind gradient-based approaches is to use training data for estimating how the prediction error depends on individual parameters of the model (which is measurable as partial derivatives of the output with respect

²For example, consider a classification problems of 2D inputs. Logistic regression is only able to *linearly separate* the inputs into two classes, i.e., to draw a line and assign all samples on one side to one class.

³Evaluation on the training set leads to overly optimistic results, especially for complex models: In the worst case, the model remembers all training set items by heart and achieves a perfect score on the training data, without any ability to generalize to new samples.

to the parameters), and then update the parameters accordingly. This is done in an iterative way until the output error becomes stagnant.

Evaluation

Evaluation is always done with respect to an *evaluation metric*, that is an algorithm that takes as input test samples, their labels and predictions for the test samples, and returns a numeric score (sometimes also referred to as *performance score*) that quantifies the overall quality of the predictions. Evaluation metrics are task-specific. For example, for classification tasks the simplest choice is accuracy, which measures the fraction of correctly classified samples. For regression tasks, metrics such as mean squared error or mean absolute error over all samples are common choices.

3.4 Adopting Supervised Learning for Subjective Interpretation

To summarize the overall set-up: In this dissertation the focus is on using machine learning models for supervised learning of subjective interpretation. Input will be a single message, output a list of concepts, concept scores or a sentence with a subjective quality.

To this end, we can generally invoke the standard supervised learning pipeline (see Section 3.3). Model selection, training and evaluation will be treated as in non-subjective cases. However, for modeling and dataset creation, extra care has to be taken when dealing with subjectivity, as will become clear in the remaining chapters.

Chapter 4

Related Work

We first give a general overview of computational approaches on subjective interpretation. After that we move to related work specific to the research questions of this dissertation.

Note that since our individual publications (Chapter 6) also contain sections on related work, in this chapter we will not go into every possible detail but focus on points that are useful for putting our work into context.

4.1 Work on Predicting Subjective Interpretation

There are mainly three application fields in computer science where subjective interpretation is central: *Information retrieval*, *data mining* and *computational communication research*.

4.1.1 Information Retrieval

The field of *information retrieval* pursues the goal of finding “relevant” contents (e.g., documents, images, videos) from a database. What exactly is considered to be relevant differs across specific retrieval tasks, but often, relevance is closely linked to personal preferences and other subjective factors.

For example, Liu et al. show that considering user history makes *search engines* more effective and efficient [76]. Large degrees of personalization in modern search engines give additional support for the fact that relevance of websites depends on the subject performing the search query. The approach described by Hanjalic et al. even delves deeper into subjectivity by explicitly considering the intent of the user when performing online searches [77].

Even before any popular web search engines existed, researchers found personalization to be useful for finding relevant technical memos [78]. The same paper mentions that information retrieval is closely related to *information filtering*. This relation is described in more detail by Belkin and Croft [79], who specify as one of the main differences between the two applications that information filtering considers long-term user interests while information retrieval focuses on short-term interests.

Another application that can be put under information retrieval is *recommender systems*, where contents (e.g., music, images, movies) are suggested to users based on their behavior. The goal thereby is to recommend contents which the user finds interesting or appealing. This can, for instance, be achieved by maintaining user profiles where preferences and interests are stored, and comparing these user attributes with attributes of contents in order to find well-aligned contents [80].

4.1.2 Data Mining

Especially in natural language processing (i.e., the branch of Artificial Intelligence that deals with natural language data), several *data mining* tasks directly focus on extracting information about subjective interpretation.

Perhaps the most common among these tasks is *sentiment analysis*, which aims at detecting the sentiment expressed in a given text. For example, given a set of movie reviews in the form of texts, one might wish to find out which movies are preferred by detecting the sentiment expressed by users in individual reviews and aggregating this information. Sentiment analysis has a long-standing tradition in natural language processing and has been extensively researched (see e.g., the surveys [81, 82]). Common methods include decision tree, SVM, neural network, Naïve Bayes and maximum entropy [82].

Aspect-based sentiment analysis is an extension of standard sentiment analysis that detects not only the overall sentiment of a text but also finds which aspect the sentiment is expressed on. For example, imagine you are given a review for a specific product. Standard sentiment analysis would detect whether the review is in favor or against the product. Aspect-based sentiment goes a step further and might find that the customer complained about the price which is claimed to be too high, i.e., there was a negative sentiment expressed toward the aspect “price”. This more fine-grained analysis adds a type of explainability that is useful for practical applications. Thet et al. [83] describe an approach based on a subjectivity lexicon and dependency parsing for extracting aspects and corresponding sentiments from movie reviews on discussion boards. Aspect-based sentiment analysis has become a popular machine learning task (see e.g., [84, 85]), hence several other approaches have been proposed, making use of topic modeling [86], Naïve Bayes [87] or neural networks (e.g., [88, 89]). There are several other noteworthy tasks which originated from sentiment analysis, including *stance detection*, *opinion mining* and *perspective detection*.

Stance detection aims at detecting the sentiment toward individual entities (people, organizations, etc.) based on textual data [90]. The main difference to aspect-based sentiment analysis is that the entity toward which the sentiment is expressed need not be mentioned explicitly in the text. Several approaches to stance detection are based on various types of neural networks, including convolutional neural networks [91], recurrent neural networks [92] and target-specific neural attention networks [93].

Opinion mining is more complex than aspect-based sentiment analysis. It aims at extracting quintuples of the form (entity,aspect,sentiment,holder,time) to capture which opinion holder expresses his or her opinion on an entity in terms of sentiment toward a particular aspect of the entity (see e.g., [81]). A concrete example for such a quintuple would be (Nokia,voice_quality,positive,user “XY”,Apr-1-2019). Extracting such information involves aspect-based sentiment detection and entity extraction as sub-tasks.

Perspective detection is a less common task but has been described by Fang et al. [94] and Vilares and He [95]. Basically, based on two text corpora (e.g., transcripts of speeches of democrats and speeches of republicans), the goal of perspective detection is to automatically describe contrastive opinions pertinent to the respective corpora. Both papers base their approach on probabilistic models. The approach proposed by Vilares and He returns a list of topics, where a topic is a list of keywords (such as “israel, iran, syria, settlement, relocation, counter-terrorism gaza, tpims, airline, metropolitan”), together with exemplary statements (e.g., “It is contrary to international law in that sense, and any nation has obligations when dealing with occupied territories and their occupants.”, ...) which summarize the opinion for each topic and corpus (examples taken from [95]).

In computer vision, sentiment analysis is not as well-established as in the text domain. Still, more recently researchers started to consider *visual sentiment* as well

[7]. Since predicting subjective visual interpretation is the topic of research question (Q1.2), below we will discuss further details (see Subsection 4.3.1).

4.1.3 Computational Communication Research

In [96], Matei and Kee describe the field of computational communication research and structure relevant efforts. Computational communication research is a diverse field and among others encompasses analysis of textual contents and user behavior.

Content analysis uses computational approaches to analyze messages in order to obtain insights about human communication. In this context, methods from natural language processing such as sentiment detection can be used for semantic analysis, but as content analysis is seen as a form of rhetorical analysis in communication research, tasks such as detecting and untangling polysemy and contextual modifiers become relevant as well.

Analysis of *user behavior* can be directly related to subjective interpretation, but this is not always the case. An example with an obvious relation to subjectivity is the analysis of likes or similar reactions that reflect user preferences and emotional assessment of contents. More complex approaches to user behavior analysis are emerging in the literature, including attempts to investigate into interactions of behavior, cognitive processes, information consumption and social networks (e.g., [97]).

Note that especially for analysis of content and user behavior, there is significant overlap with data mining. It shall also be mentioned that computational communication research is not yet as established as the more traditional fields of information retrieval and data mining. Still, as many of the approaches in computational communication research combine results from various disciplines and digital communication is now pervasive, we believe that this field will gain importance over the next years and is very interesting to consider for future research.

4.2 Modeling Interpretation

Works from philosophy, hermeneutics and psychology make clear that understanding interpretation is by no means a trivial matter and has been researched extensively. Still, these fields aim at helping humans to interpret and understand interpretation, but do not provide a general model that could directly be applied to machine learning. For example, in psychology we find Q methodology [98] which aims at describing prevalent perspectives on a given topic, but relies on human intuition for designing questionnaires and summarizing answers into coherent pictures [3]. Another example is the framework for qualitative analysis proposed by Tan et al. [4], which is based on the philosopher Paul Ricoeur’s theory of interpretation and involves complex processes such as analyzing how one’s own interpretation might be influenced by personal experiences and knowledge.

In cognitive science, many experiments are designed to collect responses which often can be seen as interpretation of stimuli. For example, in experiments on categorization subjects might be asked to rate how representative an item is for a given category, or to verify statements of the form “An [exemplar] is a [category]” (see [99]). While such experiments aim at a better understanding of human interpretation, the focus lies on identifying how certain features (e.g., some individual property of the stimuli) affect the interpretation result (e.g., perceived degree of category membership). Consequently, it is typically analyzed whether the presence of a given feature has a significant influence on the interpretation outcome, without modeling the whole interpretation process.

Closer to computer science, in the field of computational psychiatry, researchers started to test conflicting hypotheses by implementing them and comparing their

fitting ability to experimental data [100]. But as these models are developed for testing hypotheses, they are very task-specific.

However, we are looking for a general way of modeling that should be largely task-independent. Another relevant point is that models in cognitive science and computational psychiatry typically use simple input features (corresponding to controlled properties of stimuli) which renders them unfit for most applications based on multimedia messages.

So overall, in the humanities and closely related fields we do not find models that are task-independent and directly applicable to machine learning. If we turn to computer science, on the other hand, these criteria are typically satisfied, but it is rarely discussed what interpretation means. This might be surprising to theory-driven researchers, given that interpretability is currently a popular research topic in the field of Artificial Intelligence [5] and the existence of works in natural language processing on topics such as perspective detection [94, 95]. However, one has to keep in mind that research in computer science tends to be application-oriented in general. As a notable exception, the work of Montavon et al. [6] comes from computer science and defines interpretation as “mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of”. However, their definition describes interpretation in the sense of *explaining* and not as in *construing*, which we want to consider in our work (see Section 2.1). Further, their notion of interpretation does not enclose much of human interpretation, as human interpretation must start from something “that the human can make sense of” which is not necessarily abstract.

Another case worth mentioning is the work done in the field of distributional semantics. The focus there lies on learning the meaning of words, which we have found to be closely related to interpretation (see Chapter 2). Distributional semantics uses theory from cognitive science and linguistics to compute meaning in terms of co-occurring words. However, this notion of meaning is only one of several possible choices and not necessarily the best one for application, as it can cause much unnecessary overhead (see Section 2.2).

4.3 Subjective Interpretation of Images

Online, subjective interpretation of images is observable through various interactions of users with the images. These include posting titles together with images, commenting on images of others, assigning tags or reacting to the image by liking, forwarding etc. Regarding the theoretical underpinnings, as we have argued in Chapter 2, titles, tags and other reactions can be seen as purpose-specific meaning (see Subsection 2.2.5).

In the case of tags or reactions, interpreting an image then means to assign *concepts* to the image. Titles, comments or captions, on the other hand, are the outcome of a process where an image is interpreted in terms of a *phrase*. We discuss both of these forms of interpretation, starting with the comparatively simpler task of extracting subjective visual concepts.

4.3.1 Detecting Subjective Visual Concepts

We are interested in approaches for extracting informative subjective concepts from images, such that they can be useful for data mining or information retrieval. For example, a user might want to find “cute” images when in a certain mood. Or for mental healthcare it could be useful to detect an increase in sad images being posted by a certain group of users.

Even though computational approaches to subjectivity have been much more prevalent in the text domain (e.g., sentiment analysis, opinion mining, stance detec-

tion), several works have considered subjective image interpretation before. This includes efforts on detecting visual sentiment [101, 102] and image aesthetics [103, 104], or predicting image popularity [105, 106]. These approaches, however, tend to use single scalar values as output. For applications such as data mining or information retrieval, approaches with more informative output are typically preferable. Such an output format can be found in the line of work on detecting subjective adjective-noun combinations (such as “cute puppy” or “scary dog”) from images [7, 8, 107, 9, 108], based on the Visual Sentiment Ontology (VSO) [7]. The basic idea behind the VSO is to pair the subjective adjective part with a more objective noun part into concepts such as “cute puppy” or “scary dog”. These concepts which combine subjective and objective parts can directly be used for building tag-based datasets by means of crawling, and are easier to detect than mere subjective elements.

4.3.2 Subjective Image Captioning

Automatic generation of subjective image captions can be useful for making online life more convenient to users. For example, when a user uploads an image to social media, a captioning assistant could suggest a list of subjective captions the user can choose from. Other applications are entertainment (it can be quite amusing if the computer comes up with subjective texts) or explaining image contents to visually impaired people (who can be assumed to also have some interest in the subjective aspects of posted contents).

Originally, image captioning was restricted to generation of objective descriptions. In early works the task was treated as a retrieval problem, where in order to find a suitable caption for an image, cases of similar images with known captions were retrieved from a large dataset and their captions were suggested [109]. Nowadays, the most common approaches are based on the neural network-based method “show-and-tell” [110]. In this method, the image is first converted into a vector by one neural network, then this vector is transformed into a caption by another neural network.

Another approach to image captioning is to split the task into two steps, where the first step is to extract concepts and the second step is to turn the extracted concepts into a caption using a language module. The language module is typically a simple statistical model, where the type of concepts that are extracted in the first step varies. For example, Leuret et al. [111] extract phrases such as “a skate board” or “is riding”, while Fang et al. [112] detects individual words as concepts and Li et al. [113] use n-grams.

Less work has been done on subjective image captioning. While there exists some work done on personalization of captions, for example by adding face recognition to the processing pipeline (e.g., [114, 115]), only few results about generating affective captions have been reported. To the best of our knowledge, the only paper on affective image caption generation that was published before our work on subjective image captioning is the one by Mathews et al. [116], where an end-to-end method called SentiCap was introduced. However, SentiCap requires specific data to train (captions with strong sentiment), and since it only outputs a final caption for a given image, it can be difficult to interpret failure cases and debug the model.

4.4 Gang Violence Prevention

There are efforts by US law enforcement to address gang violence. For example, the city of Chicago compiled a “strategic subject list” [117], predicting who is likely to shoot someone or be shot, with the goal of having police officers talk to these people preventively. But for their approach they rely on non-public offline data, and the actual benefit for the community has been severely criticized [118, 119].

As we mentioned above in the introduction, social scientists analyzed online behavior of gang members and found them to be as active as “normal” youth on various social media platforms [15], including public posting. Another relevant observation is that online violence (“cyberbullying”) can lead to offline crime [23, 24]. These two observations can be combined into the idea of detecting problems online before they turn into physical conflicts. In the literature, we find a few works on using social media posts for fighting crime. For example, Gerber applies statistical topic modeling to tweets with geolocation for predicting how likely 20 different types of crimes are to happen in individual parts across the city of Chicago [120]. However, this work is a large scale statistical approach that is not specific to gang associates and is meant to help placing police officers in an efficient way, whereas our efforts are community based and have solid grounding in social work research.

Most closely related to our work is the paper by Blevins et al. [10]. They detect *aggression* and *loss* from tweets of a deceased gang member and her top communicators, using an extensive set of linguistic features. In fact, this work was done at Columbia University and involved the Natural Language Processing group and SAFELab, both of which also took part in the efforts presented as part of this dissertation.

Chapter 5

Own Work

In this chapter, we explain which approaches we adopt for each of the three research questions, and how these efforts relate to the publications included in the next chapter.

5.1 Modeling Interpretation

In our survey paper “*An Overview of Computational Approaches for Interpretation Analysis*” (pages 39ff.) we formally define interpretation for machine learning in terms of an interpretation function of someone/something that maps from input to meaning, and describe the task of computational interpretation analysis. There are two crucial differences as compared to related definitions of interpretation in computer science (e.g., [6]): First, we do not pose any a priori restrictions on the input and output domains. This makes it applicable to human as well as machine interpretation. Second, we introduce a bearer of the perspective. This modification was inspired by the treatment on points of view by Campos and Gutiérrez [1], and proves helpful for comparing multiple ways of interpretation.

We see that, mathematically speaking, analysis of interpretation essentially means to characterize functions, which can generally be done by describing dependencies between the function’s input and output. This simple insight helps to find many approaches for analyzing and comparing ways of interpretation. Consequently, in the same paper we provide a comprehensive overview of relevant approaches for analyzing interpretation (including statistical methods, pattern mining, model-based approaches, visualization techniques) and explain methods for comparing multiple ways of interpretation with computational methods.

Furthermore, our framework for interpretation analysis is analogous to the standard supervised learning formulation, which means that analysis and prediction can both be treated in this same framework. In particular, we describe the possibility of using a single neural network for prediction and analysis of multiple ways of interpretation. Such a neural network is trained to predict an interpretation (e.g., like vs dislike) from input x (e.g., image) and perspectival context p (e.g., ID of user expressing the preference). Once trained, the network can be analyzed in order to derive insights about properties of the learned interpretation functions (e.g., which users share similar preferences). This possibility leads to two open questions: First, which architecture to use for merging input and context information? And second, how reliable is the information about the perspectives that is learned by the network? Put together, these open questions motivate the next publication.

In “Fusion Strategies for Learning User Embeddings with Neural Networks” (pages 95ff.), we train a neural network to predict movie ratings from movie features (input) and user ID (context), and analyze how the way of merging input and context infor-

mation influences what the network learns about the users’ ways of rating movies. We also see how understanding interpretation as a function reveals the relevance of the mathematical field *Functional Data Analysis*, which shows how to properly measure distances between ways of interpretation. We used this understanding to introduce a metric for evaluating the quality of vector representations of interpretation functions.

5.2 Subjective Image Interpretation

5.2.1 Detecting Subjective Visual Concepts

Methods for detecting adjective-noun pairs based on VSO [7] (e.g., [8]) have led to a great improvement in the ability to detect subjective concepts from images. However, VSO suffers from several shortcomings. First, the different adjective-noun combinations are assumed to be independent. This is of course an inaccurate simplification, since for example the interpretation “adorable puppy” would normally imply “cute dog” but exclude “scary dog”. During evaluation based on VSO (or its extension MVSO [107]), this simplification becomes problematic, as “cute dog” would generally be considered wrong for an image labeled as “adorable puppy”, making it more difficult to judge the quality of detection models. Second, as much as the trick of combining adjectives and nouns helps for easier detection of subjective concepts, it simultaneously waters down the subjective component. Thus, from reported performances for detecting adjective-noun combinations it is not directly clear whether the model mostly specializes on distinguishing the different possible nouns or focuses on adjectives.

We address these shortcoming by introducing the Focus-Aspect-Value (FAV) model for subjective image interpretation in our paper “*The Focus-Aspect-Value Model for Explainable Prediction of Subjective Visual Interpretation*” (pages 103ff.). In the introduced model, we assume that a focus in the image is determined by selecting a noun. The task then is to find out which adjectives make sense or are likely to be suitable for describing the part of the image that is in focus. To this end, adjectives are organized into mutually exclusive sets. Based on Google’s Conceptual Captions dataset, we created a new dataset *aspects-DB* following the FAV model and evaluate various machine learning methods on this new task.

As mentioned earlier, detecting subjective visual concepts can be useful for image retrieval or data mining. Yet another application which can benefit from such work is subjective image captioning.

5.2.2 Subjective Image Captioning

For subjective image captioning, we have three relevant publications: First, in “*Image Captioning in the Wild*” (pages 113ff.) we describe a crowdsourcing study on how people interpret images on Flickr with image captions. In particular, in this study we found that there are only few captions with fully visible and purely objective contents, and annotators in general preferred captions with sentiment.

Both observations justify a shift toward more subjective image captioning. We worked on this task by adopting the pipeline approach outlined in Figure 5.1: In our approach, we first detect subjective concepts (such as adjective-noun combinations) and then use a language model to turn them into a phrase. This has the advantage that the language model can be built separately under the use of any text-only corpus, which renders it unnecessary to have any subjective captions as ground truth. Also, the detected subjective concepts can be used for additional explanation and verification of the model.

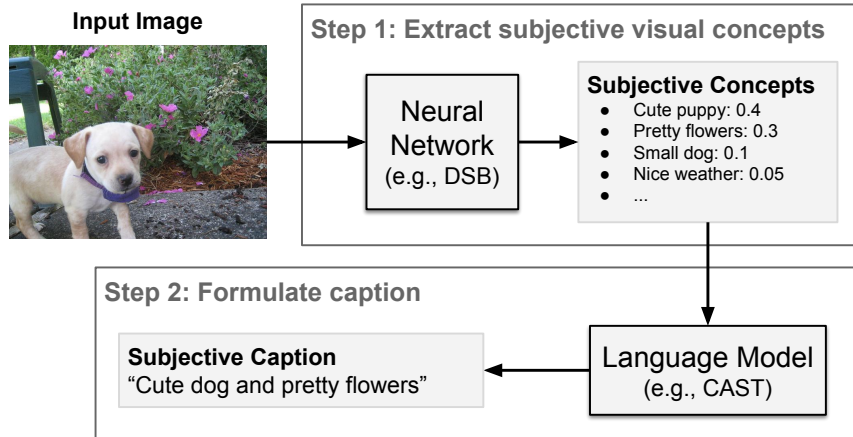


Figure 5.1: Our overall approach to subjective image captioning. The first step is to use a subjective visual concept detector, then a language generation module is used to turn these into a subjective sentence. The input image is taken from Flickr, where it was published by DennisAHansen with the title “IMG_1084” (<https://www.flickr.com/photos/dennishansen/32740334477/>) under Public Domain Mark 1.0.

This approach is described in two papers “*Introducing Concept and Syntax Transition Networks for Image Captioning*” (pages 123ff.) and “*Generating Affective Captions Using Concept And Syntax Transition Networks*” (pages 127ff.). Both papers use the neural network DeepSentiBank [8] for extracting subjective visual concepts, followed by a novel language generation module we call CAST. The main difference between our two image captioning papers can be found in evaluation: In one paper, the resulting captions are evaluated by human judges according to how natural they are (in a set-up similar to the Turing test); in the other paper evaluation is done mainly with respect to appropriateness. Finally, we shall mention that both modules in our pipeline approach can in principle be replaced. For example a detector based on the FAV model can take the place of DeepSentiBank for subjective visual concept detection, and a recurrent neural network could be used instead of CAST as language generation module.

5.3 Gang Violence Prevention

We extended the work of Blevins et al. [10] in several ways, most importantly by also considering visual information posted in the tweets. This decision was motivated by the fact that many tweets include images and the observation that such images can contain important information for intervention workers. For example, images might show which people hang out together, which kinds of firearms these groups possess or which kinds of substances they abuse. Also, for some posts with images, aggression is visible only when considering the image (e.g., when a gun is pointed at the camera with an ambiguous text).

In Figure 5.2 we illustrate our overall approach to gang violence prevention. For preventing online aggression from escalating to offline violence, we aim at detecting problems online such that local intervention workers can be informed and approach specific individuals before the situation escalates. This approach involves the three steps analysis, notification and intervention. We only handle the first of these steps in this dissertation, because it most crucially involves automatic prediction of interpretation.

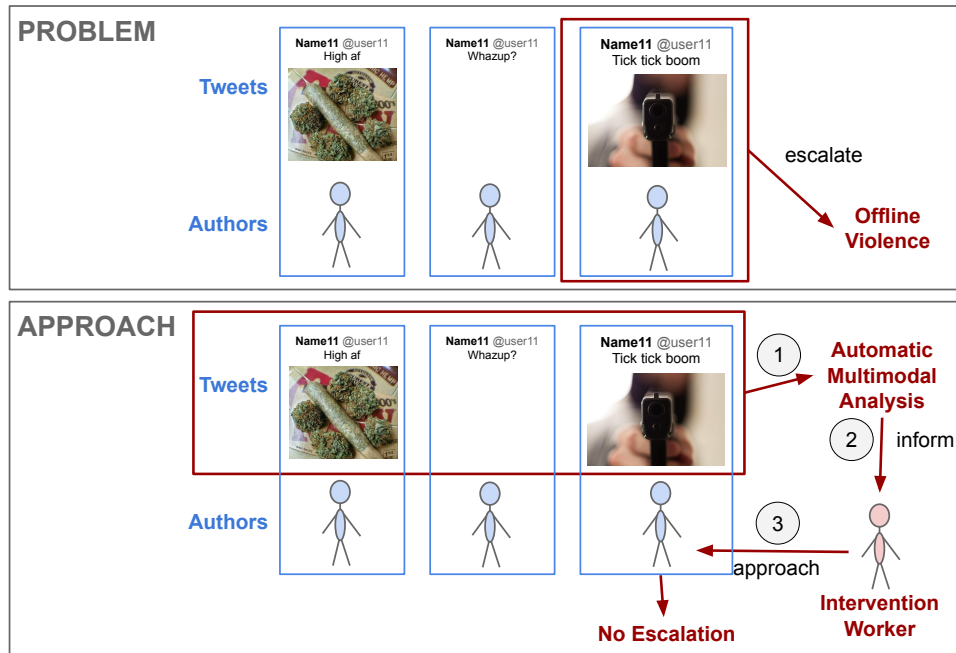


Figure 5.2: Illustration of our approach to gang violence prevention. Gang-involved youth are active on social media platforms such as Twitter, where some of their posts lead to offline violence. Our approach to prevent violence is to automatically analyze social media posts with the goal of extracting information that is useful for local intervention workers for taking preventative measures. The focus of this dissertation is on automatic tweet analysis. Images are taken from Flickr for illustration purposes: “Weed Marijuana Cannabis 420” by Weed Streetwear (<https://www.flickr.com/photos/149531682@N02/33870793081/>), released under Public Domain Mark 1.0, and “The Robbery” by Geoffrey Fairchild (<https://www.flickr.com/photos/gcfairch/4189169360/>), released under CC-BY-2.0.

Note that, unfortunately, when aggression is detected it is often too late to interfere which makes the problem more difficult. In order to address this problem, we are seeking to not only detect aggression but find early indicators as well (e.g., expression of loss was found to indicate aggression around a week after), and extract other information potentially useful for informing intervention workers or developing a better understanding of how escalation builds up. (This part is not illustrated in the figure in order to keep the idea simple.) This was also one of the reasons for adding the code *substance use* as one of the tweet labels, which constitutes another difference to the work of Blevins et al.

In our paper “*Multimodal Social Media Analysis for Gang Violence Prevention*” (pages 133ff.) we describe the full process of building a multimodal analysis system for gang violence prevention. To this end we created a new dataset consisting of (public) tweets with images from presumably gang-associated youth, together with annotations for the codes *aggression*, *loss*, *substance use*, plus 9 local visual concepts. For collecting these annotations, we developed the open-source system VATAS for annotating social media data, which we released with the paper “*VATAS: An Open-Source Web Platform for Visual and Textual Analysis of Social Media*” (pages 145ff.).

Importantly, in this application study we are dealing with a marginalized and vulnerable community. In addition, annotation required special efforts due to subjectivity of the involved concepts and potentially detrimental consequences of incorrect

labeling. This motivated us to collaborate closely with social work researchers for all of this work (our paper on VATAS describes at length our interdisciplinary approach between social work research and computer science for annotation), and to look closer into annotation disagreements with the paper “*Annotating Twitter Data from Vulnerable Populations*” (pages 171ff.). In this paper, we introduce new methods for explaining disagreements between annotators. We propose qualitative and quantitative methods for doing so, where I was developing the quantitative methods.

Chapter 6

Publications

This chapter contains all publications of this dissertation:

Modeling Interpretation (Q1.1)

- Section 6.1 (pages 39ff.):

An Overview of Computational Approaches for Interpretation Analysis.

P Blandfort, DU Patton, J Hees. (submitted in 2018, revised in May 2019). *arXiv preprint, arXiv:1811.04028v2*. arXiv. URL: <https://arxiv.org/abs/1811.04028v2>

- Section 6.2 (pages 95ff.):

Fusion Strategies for Learning User Embeddings with Neural Networks.

P Blandfort, T Karayil, F Raue, J Hees, A Dengel. (2019). In *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN 2019), July 14–19, 2019, Budapest, Hungary*. IEEE, Piscataway, NJ, USA. DOI: <https://doi.org/10.1109/IJCNN.2019.8852259>

Subjective Interpretation of Images (Q1.2)

- Section 6.3 (pages 103ff.):

The Focus-Aspect-Value Model for Explainable Prediction of Subjective Visual Interpretation.

T Karayil*, **P Blandfort***, J Hees, A Dengel. (2019). In *2019 International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada*. ACM, New York, NY, USA, Article 4, 9 pages. DOI: <https://doi.org/10.1145/3323873.3325026>

- Section 6.4 (pages 113ff.):

Image Captioning in the Wild: How People Caption Images on Flickr.

P Blandfort*, T Karayil*, D Borth, A Dengel. (2017). In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes (MUSA2 '17)*. ACM, New York, NY, USA. DOI: <https://doi.org/10.1145/3132515.3132522>

- Section 6.5 (pages 123ff.):

Introducing Concept and Syntax Transition Networks for Image Captioning.

P Blandfort*, T Karayil*, D Borth, A Dengel. (2016). In *2016 International Conference on Multimedia Retrieval (ICMR '16), June 06–09, 2016, New York, NY, USA*. ACM, New York, NY, USA. Pages 385–388. DOI: <https://doi.org/10.1145/2911996.2930060>

- Section 6.6 (pages 127ff.):

Generating Affective Captions using Concept And Syntax Transition Networks.

T Karayil*, **P Blandfort***, D Borth, A Dengel. (2016). In *Proceedings of the 24th ACM International Conference on Multimedia (MM '16'). October 2016, Amsterdam, Netherlands*. ACM, New York, NY, USA. Pages 1111–1115. DOI: <https://doi.org/10.1145/2964284.2984070>

Gang Violence Prevention (Q1.3)

- Section 6.7 (pages 133ff.):

Multimodal Social Media Analysis for Gang Violence Prevention.

P Blandfort, DU Patton, WR Frey, S Karaman, S Bhargava, FT Lee, S Varia, C Kedzie, MB Gaskell, R Schifanella, K McKeown, SF Chang. (2019). In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2019), 13(01), 114-124*. AAAI, Palo Alto, CA, USA. Retrieved from <https://www.aaai.org/ojs/index.php/ICWSM/article/view/3214>.

- Section 6.8 (pages 145ff.):

VATAS: An Open-Source Web Platform for Visual and Textual Analysis of Social Media.

DU Patton*, **P Blandfort***, WR Frey, R Schifanella, K McGregor, SF Chang. (2020). *Journal of the Society for Social Work and Research 11(1)*. The University of Chicago Press, Chicago, IL, USA. DOI: <https://doi.org/10.1086/707667>

- Section 6.9 (pages 171ff.):

Annotating Twitter Data From Vulnerable Populations: Evaluating Disagreement Between Domain Experts and Graduate Student Annotators.

DU Patton, **P Blandfort**, WR Frey, MB Gaskell, S Karaman. (2019). In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. University of Hawai'i at Manoa, Honolulu, HI, USA. URI: <https://hdl.handle.net/10125/59653>

*Denotes equal contribution.

All publications are included exactly as they are published (or as they were accepted in case of the second paper), except that a header was added in order to ease navigation and integrate the contents into this dissertation. In particular, the papers appear in their original typesetting and some of them show additional page numbers at the bottom which should be ignored.

An Overview of Computational Approaches for Interpretation Analysis

Philipp Blandfort

TUK and DFKI, Kaiserslautern, Germany

Jörn Hees

DFKI, Kaiserslautern, Germany

Desmond U. Patton

Columbia University, NYC, USA

Abstract

It is said that beauty is in the eye of the beholder. But how exactly can we characterize such discrepancies in interpretation? For example, are there any specific features of an image that make person A regard an image as beautiful while person B finds the same image displeasing? Such questions ultimately aim at explaining our individual ways of interpretation, an intention that has been of fundamental importance to the social sciences from the beginning. More recently, advances in computer science brought up two related questions: First, can computational tools be adopted for analyzing ways of interpretation? Second, what if the “beholder” is a computer model, i.e., how can we explain a computer model’s point of view? Numerous efforts have been made regarding both of these points, while many existing approaches focus on particular aspects and are still rather disconnected.

With this paper, in order to connect these approaches we introduce a theoretical framework for analyzing interpretation, which is applicable to interpretation of both human beings and computer models. We give an overview of relevant computational approaches from various fields, and discuss the most common and promising application areas. The focus of this paper lies on interpretation of text and image data, while many of the presented approaches are applicable to other types of data as well.

Keywords: survey, interpretation, analysis, perspective, explainability,

machine learning, pattern mining, visualization, correlation, social science

1. Introduction

Individual ways of interpretation play a major role in a variety of fields. The philosophical positions scepticism, relativism and perspectivism all crucially involve the notion of points of view [1], i.e., different ways of interpretation. Hermeneutics refers to a whole field that is concerned with how we interpret information and commonly assumes that in order to make sense of things we need to relate them to our own life situation, which makes all interpretation something inherently personal (e.g., see [2]). Analyzing how we make sense of the world is pertinent to cognitive science, the research field concerned with studying the human mind. Similarly, in psychology it has been argued that understanding each others' motivations is a key aspect of human social life [3]. Even in a non-scientific context, everyday misunderstandings in communication offer a clear demonstration of both challenge and importance of correctly estimating what other people mean and anticipating how they would interpret our own behavior.

Nowadays, there are two developments that drastically impact our social life and motivate the need for computational methods with similar social abilities: First, more and more communication is happening online [4]. Second, AI approaches have become much more ubiquitous. This is especially prevalent online, where chatbots take part in discussions, recommendation algorithms suggest things we are likely to favor, and search results are nicely ranked by yet another computer model. In a broad sense, humans and computer models are all actors in a large communication network. In many cases the goal of an AI approach is to learn about a certain way of interpretation. This is most clearly visible in supervised approaches where the ground truth data serves as a proxy to the human perspective that is to be learned, which often involves estimating subjective qualities (e.g., what a user will like, or even automatically mining opinions). At the same time, as AI approaches become actors in communication and their automatic decisions become more and more influential in our everyday life, we also have a motivation to understand them. As approaches have grown considerably more complex over the years, this is not at all trivial. However, since early 2018, with changes in European legislation (GDPR [5]) there is now even a legal reason why many companies (and probably also researchers) should analyze how the developed

models draw their conclusions: Whenever users are affected by automatic decisions, the users now have the legal right for an explanation of the decision in simple terms [6]. Yet another pragmatic motivation for understanding AI approaches stems from ever-growing amounts of data (“big data”) involved in digital activities such as posting comments, liking contents or browsing websites: Due to the scale of user data, it has become extremely challenging to manually inspect even a fraction of the data. Here, computers have a clear edge in terms of scalability, and are valuable for processing all this information and thus making it more accessible to us, potentially even by explaining its characteristics.

So we see that there are three important tasks, namely enabling AI approaches to “understand” our view, understanding how AI agents see the world, and having computer models explain complex data to us. It is clear that neither of these tasks is simple, still, good progress has been made on all of them. To name a few recent advances: A lot of work was done on explaining how deep learning models work [7, 8, 9, 10, 11, 12, 13], which was even useful for helping us understand complex scientific data [14, 15, 16]. In case of data annotation, probabilistic methods have been proposed to merge annotator votes efficiently and simultaneously estimating annotator reliabilities [17, 18]. However, despite related goals, approaches for interpretation analysis seem quite separated and we find an apparent lack of high-level bridges to connect them. In particular, recent surveys on explainability methods for machine learning [7, 8, 9, 10] do not consider methods for comparing multiple ways of interpretation. Moreover, underlying concepts such as interpretation or understanding are often not defined properly (as [10] explains for the concept interpretability), which suggests the need for more rigorous formalism.

The main purpose of this paper is to connect various ideas and approaches, and put them into a coherent view. To this end, we introduce a theoretical framework, in which a perspective is represented by a function from input to meaning, called the interpretation function. Interpretation analysis can then be understood mathematically as characterization of such an interpretation function. We do a survey on approaches for this task with a focus on text and image inputs, where we in particular find statistical methods, pattern mining, model-based approaches and visualization techniques to be of central relevance. In addition to outlining methods for analyzing interpretations of a single model, this paper describes methods for comparing multiple perspectives. We also unveil relations to the humanities, where it

has a much longer tradition to look into characteristics of interpretation, in the hope that this will contribute to more discussion between the disciplines.

We structure the paper as follows: First, in Section 2 we will describe our theoretical framework and formally define interpretation, perspective and the task of interpretation analysis. This is followed by general remarks about the task in Section 3, where we comment on evaluation, ethics and input representation. We will then look into approaches for the case of analyzing one individual perspective (Section 4). To this end, we can make use of statistical methods, pattern mining, model-based approaches or visualization techniques (see overview in Table 1). Comparisons between multiple perspectives will be handled in Section 5 and can be done under the use of three kinds of approaches (see also Figure 3). We will see that two of these cases can mostly be reduced to single perspective analysis, which makes the methods for analyzing relations between input and output of a single interpretation function the core of this paper. In Section 6, we outline five application fields, where ways of interpretation are analyzed by means of computational methods. Finally, we close the paper with a few remarks on future work and ethical aspects (Section 7).

2. Theoretical Framework

Montavon et al. [9] define interpretation as a “mapping of an abstract concept (e.g., a predicted class) into a domain that the human can make sense of”. We agree that this might work for the specific purpose of their analysis, but find this definition to be in conflict with intuition. Most importantly, the definition does not include a large part of human interpretation, which in general starts from something concrete (like an image or text) and ends up in something more abstract that we can broadly call meaning. Hence, we keep the mapping part but remove the restrictions of the input and output domain while we introduce the notion of a *bearer*, inspired by recent works in philosophy on defining perspectives [1, 19]:

Definition (Perspective, bearer, interpretation, interpretable)

We define a *perspective* as a way of interpretation of some actor or group of actors b , which we call the *bearer(s)* of the perspective. Formally, we can represent a way of interpretation by a mapping from input to meaning, and call this mapping the *interpretation function* f_b of b :

$$f_b : I_b \rightarrow M_b, \tag{1}$$

where I_b is the input domain and M_b the output domain (set of potential meanings). Any information i is then called *interpretable* by b if and only if it is contained in the input domain of b 's interpretation function, i.e., $i \in I_b$.

Examples

1) Image classification of a machine learning model m can be seen as interpretation process, where the interpretation function f_m of the model maps from a set of images I_m into a set of classes M_m . 2) An example for a human perspective would be the interpretation process of annotator a from a set of tweets I_a into {sarcastic, not sarcastic} when being asked to label tweets accordingly. 3) More complicated output domains are possible. For example, in case of an image autoencoder e the latent representation can be modelled as interpretation of e .

2.1. Role of the bearer

We do not impose any particular requirements on the input or output domain, but we require that a perspective is adopted by some actor b (e.g., human being or computer model, existing or hypothetical), or group of actors. In case a restriction is necessary, one can achieve this by limiting the set of possible bearers, which naturally leads to restrictions on the input and output domains, as well as the form of possible interpretation functions. For example, if b is limited to be certain neural networks, both inputs and outputs are typically in tensor format.

Introducing a bearer in our definition of interpretation also paves the way to comparing ways of interpretation adopted by different bearers. This can mean comparing perspectives of different people or perspectives of a single person under different circumstances (e.g., happy vs sad). In this way, our theoretic framework can be used to analyze the effects of contextual factors such as mood, geolocation or preceding events on interpretation.

Note that in the following, if only a single perspective is involved, we will usually not explicitly mention the bearer of the perspective and just use the symbol f to refer to the interpretation function.

2.2. Assumptions

For this paper, we assume that we do not have direct access to any interpretation function f , but only have a list of inputs and their corresponding outputs. In other words, we treat interpretation as a black-box, that is accessible only through a list of input-output pairs. More precisely, if a single

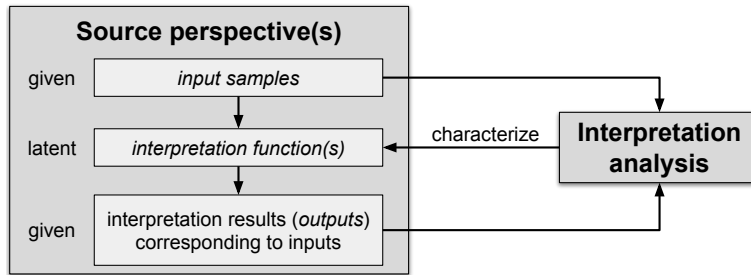


Figure 1: Interpretation analysis under the black-box assumption. The goal is to characterize interpretation from one or several perspectives, which can be human or artificial. Interpretation from each perspective is formally described as a mapping from information to meaning. For this paper, we assume that these functions are not directly accessible, but only indirectly via a list of inputs and associated outputs.

perspective is analyzed, the data is of the form $(d_0, f(d_0)), \dots, (d_n, f(d_n))$, $n \in \mathbb{N}$. Analogously, if multiple perspectives are involved, we assume the data to be of the form $(d_0, b_0, f_{b_0}(d_0)), \dots, (d_n, b_n, f_{b_n}(d_n))$, where b_i describe the bearers of the respective perspectives.

The assumption that interpretation functions are not directly observable and perspectives are given indirectly as input-output pairs enables us to more easily model interpretation of humans and AI approaches within the same framework. This is another point that clearly distinguishes this survey from other overview papers related to explainability such as [10, 8, 9], which assume that f stems from a known machine learning model.

2.3. Goals of interpretation analysis

Overall, the main goal of interpretation analysis is to characterize interpretation functions. (See Figure 1 for a schematic overview.) Such a characterization can take different forms and be addressed in various ways, depending in particular on whether the goal is to understand a single perspective (Section 4) or to compare several perspectives (Section 5).

For analysis of a single perspective, we want to extract characteristic properties from a single function in order to answer the question: “What are the relations between features of the input and interpretation result?” For example, which parts of the image make the classifier say that there is a dog in the image?

Approach type	Methods	Outcomes
statistical methods	correlation coefficients; hypothesis testing; CCA	measure of correlation, significance, canonical correlations
pattern mining	association rule mining; emerging pattern mining; discriminative pattern mining	association rules (implications), characteristic patterns
model-based approaches	heatmapping; prototypes; globally understandable models; partially understandable models; ablation studies	model for approximating interpretation function, plus: explanations for individual decisions (heatmapping), characteristic inputs (prototypes), or approximate functional description of the function
visualization techniques	dimensionality reduction; example-based approaches; text summarization	compression of the data, in form of plots, selected examples, or text summary

Table 1: Overview of approaches for single perspective analysis.

For comparing several perspectives, we are generally interested in discriminative characterization. For example, we can ask “For which kinds of inputs can we expect any difference between machine learning models A and B?” or “Which features of tweets characterize the set of tweets which annotator C labels as *aggressive* while annotator D labels them as *non aggressive*?”

3. Computational Approaches

As we just saw in Section 2.3, interpretation analysis in the proposed framework amounts to characterizing functions, interpretation functions to be more precise. The general purpose of functions is to formally describe how one quantity (the output) depends on another quantity (the input). Hence, at the very core of interpretation analysis (or analyzing and understanding any function for that matter) we find the task of figuring out how outputs *depend on* inputs. And this is to be done based on a list of inputs

and their corresponding outputs. So we have already converted the conceptually challenging problem of interpretation analysis into a more graspable mathematical formulation, which can be tackled with a variety of existing computational methods. We have also discussed that the task takes on a slightly different touch depending on whether we are analyzing one individual perspective or aim at comparing between multiple ones. Before we go into detail on these approaches in Sections 4 and 5, we will first discuss three general points that are relevant in all these cases, namely evaluation, ethics and feature extraction.

3.1. Evaluation

Natural questions to ask when being confronted with any large set of tools for a single task are: Which one to choose? And on which grounds should one make such a decision? So, how can we evaluate which method for interpretation analysis does the best job?

First of all, despite following the common goal of characterizing a single function in terms of relations between input and output, the relevant approaches vary in terms of result format, but also with respect to other properties such as reliability and expected data (type and amount). This makes it difficult to directly compare all the approaches, and indeed, a general automatic evaluation measure for interpretation analysis does not exist. For several individual categories evaluative measures have been proposed (e.g., see [20] for heatmapping), but in practice, quantifying usefulness of explanations largely remains an open issue and qualitative evaluation often becomes necessary. This can mean that researchers manually inspect results and view examples for judging which method does the better job, or task someone else (e.g., crowdworkers) with evaluating which method generates better explanations (e.g., as in [21]). Another interesting option is mentioned in [9], namely to look at simpler versions of the tasks where an optimal explanation can be specified and then compare the results to this explanation.

In general, we regard the following three criteria as important: 1) The results should be *reliable*, which includes statistical significance and robustness. 2) The characterization should be simple to *understand*. 3) The findings should *cover* as much as possible of the *variation* in the data that one wants to understand. (For a single perspective, explain variations in output in terms of input; for several perspectives, explain their differences.) Note that these points are treated quite differently in the relevant fields. Reliability is absolutely fundamental in statistics and still important in pattern

mining, but mentioned more rarely in model-based approaches. Understandability is a factor across the fields, but interestingly, the necessary background knowledge for correctly interpreting given explanations varies significantly. Coverage of variation is often checked in statistics (coefficient of determination, R^2), quite central in pattern mining, but harder to address in some of the model-based approaches (e.g., how to measure to which degree output variation can be explained in terms of heatmaps or prototypes).

3.2. Ethics

We have just discussed various general criteria for judging the quality of analysis methods given a specific task. However, if we zoom out and look at the big picture of interpretation analysis, it becomes clear that such analyses often have substantial ethical implications. Thus, we find ethical considerations to be a crucial part of analysis, especially when dealing with human interpretation.

Analyzing human interpretation

Research on human interpretation can help to improve user experiences but also pave the way to ethically doubtful applications. For example, better understanding how we interpret information can be used for “computational propaganda” [22] and microtargeting, where people’s personality traits are used to predict what kind of message is most likely to persuade them [23].

A nice starting point for ethical considerations can be the paper by Zook et al. [24], which introduces “ten simple rules for responsible big data research”, including many examples and pointers to further details. We cite their ten rules here to provide a general idea, while we refer to their paper for details:

1. “Acknowledge that data are people and can do harm”
2. “Recognize that privacy is more than a binary value”
3. “Guard against the reidentification of your data”
4. “Practice ethical data sharing”
5. “Consider the strengths and limitations of your data; big does not automatically mean better”
6. “Debate the tough, ethical choices”
7. “Develop a code of conduct for your organization, research community, or industry”
8. “Design your data and systems for auditability”

9. “Engage with the broader consequences of data and analysis practices”
10. “Know when to break these rules”

Importantly, these points should encourage thinking and discussing about ethical implications in the first place, but also make clear that ethics is not a simple matter. In this context, we would like to recommend to not only discuss with researchers from computer science but form interdisciplinary collaborations. This certainly does not automatically eliminate all potential negative consequences, but we believe that it does reduce the risk by safeguarding against very narrow perspectives. Overall, we advice to start by asking questions such as “Do we really want to analyze this aspect of interpretation?” and “Could such an analysis potentially do more harm than good?” before jumping into technical details.

Interpretation analysis in the broader context of AI

Recent advances in AI suggest a great potential for solving pressing social problems with the help of computer systems, while building ethical AI requires us to wrestle with tough questions like “Is this moral?”, “Is it racist?”, “Is it safe for everyone?”, “Should we build it?”

Take for example the issue of predictive policing. There is a growing trend among law enforcement units globally (cities like Chicago, London and New York City) where big data and machine learning are used to predict potential criminals and surveil communication on social media platforms [25]. Early on, this form of digital policing was touted as an innovative strategy for catching crime and violence before it happens [26]. However, researchers and journalists have identified clear challenges that include: unconscious and implicit bias in the interpretation of language and images on social media that are deemed threatening [27], increased and disproportional surveillance of black and brown communities [28], increased arrest of individuals who pose little threat and missed predications of white perpetrators of crime and violence [29].

One practical response is the creation of critical and diverse partnerships between computer scientists, community members and law enforcement that reviews interpretation of images and text across race, ethnicity and culture, analyzes system outputs for racial and cultural sensitivity, and considers the implications of AI tools for community well-being and safety. Within such an environment, we see great potential for interpretation analysis techniques by using them for revealing problematic biases in training data or AI systems.

3.3. Feature extraction

Lastly, the third generally applicable point is feature extraction. Here, and in most of machine learning, we face a situation similar to that in correlational studies in psychology [30], where the data is already there and we need to answer: What is the kind of input “parts” we want to consider for checking dependencies with the output?

First of all, many of the approaches we will discuss cannot be expected to reveal interesting findings when applied to low-level input features such as individual pixels or sequences of characters. For example, if the color of any individual pixel of an image correlates significantly with a classifier output for “dog”, then this is hard to make sense of and has a high chance of being a statistical artifact or a flaw in the training data. This is per se not specific to interpretation analysis and especially in applied machine learning feature engineering (i.e., finding suitable features) remains a key part [31] despite the efforts of the deep learning community for end to end learning. This process generally requires expertise, since the features need to be appropriate for the final method, the data at hand, and the overall purpose of analysis. It is in the last of these parts, purpose of analysis, where we find a considerable difference between standard machine learning and interpretation analysis. Most of the time, in machine learning the features are meant to serve the purpose of building a prediction model that is reliable (i.e., does not overfit) and has good predictive power. In case of interpretation analysis, we have seen both of these criteria in similar forms (predictive power corresponding to coverage of variance), but in addition require that results should be understandable (see Section 3.1).

This leads to some features such as intermediate activations of a Convolutional Neural Network (CNN) being less straightforward to use. After all, if for instance the 10th neuron of the penultimate layer from a VGG network [32] was found to correlate with another image classifier’s positive decision for the dog class, wouldn’t this tell us more about VGG-based embeddings than about how the classifier interprets images?

Still, when deciding on which features to use, one should definitely be inspired by existing approaches on feature extraction, and some of the simpler common features (e.g., bag of words, occurrences of specific n-grams, color histograms, bag of visual words) can be useful for analyzing interpretation. Finally, in interpretation analysis it happens at times that features are implied by the research goal. For example, if one wants to analyze whether a visual sentiment classifier prefers cats over dogs, cat and dog presence are

suitable features. Overall, finding the right features is a complex topic, in part because the understandability criterion is hard to formalize and its implications depend on the type of approach that is used later. Hence, we will mention approach-specific examples in some of the following sections (4.2 and 4.3).

4. Input-output dependencies

We now discuss computational approaches for understanding a single perspective. Typical examples would be to analyze which words in a social media post correlate with large numbers of likes (as a form of positive interpretation from the group of viewers), or to analyze an image classifier based on a list of image-classification results for identifying which patches of images are most relevant for a particular result.

The formal context can be summed up as follows (also see Section 2): The perspective is described by an interpretation function $f : I \rightarrow M$ of interest. This function is not given directly, so the goal of analysis is to determine relations between the function's input and output based on a list of input-output pairs $(d_0, f(d_0)), (d_1, f(d_1)), \dots, (d_n, f(d_n))$, where $n \in \mathbb{N}$ and $d_i \in I$ for all i . We are primarily interested in cases where I consists of language data, images, or feature vectors thereof. The output domain M is assumed to contain feature vectors of fixed dimension.

For such a task we have several types of approaches from various well-established fields at our disposal, which we will now discuss. We group these approaches together into sections which roughly correspond to research fields (statistical methods, pattern mining, model-based, visualization). In each section, we then organize techniques by their outcome or the goals they are aiming at. Each section is concluded with remarks on the usage of the respective type of analysis approach. (See overview in Table 1.) For several of these, we will use hypothetical user preference data for illustration. This data can be found in Table 2 and corresponds to a simple interpretation from a 3-D feature space into the binary space of like/dislike.

4.1. Statistical methods

One way to analyze relations between two quantities is to test for statistical dependencies between them. We can treat both input and output as values of (composed) random variables X and Y respectively, and then

Image ID	Nudity	Humor	Explosions	Like
0	0	1	0	0
1	1	0	1	1
2	0	1	1	1
3	1	1	0	0
4	1	0	0	1
5	1	1	1	0

Table 2: Hypothetical image preference data of a single user. The three columns in the middle describe features of the image, while the last column describes the type of user reaction which corresponds to an interpretation result (assuming the user has the option to e.g., either vote up or not).

test whether individual dimensions X_i of X and Y_j of Y are *statistically dependent*. Formally, such a dependency is given if for any sets of possible values A and B , the two events $X_i \in A$ and $Y_j \in B$ are not independent, i.e., $P(Y_j \in B \mid X_i \in A) \neq P(Y_j \in B)$. In other words this means that information about the value of X_i can give us any information about the value of Y_j . In our toy example (Table 2), we could check if the image preference of the user statistically depends on whether the image contains nudity, humor or explosions. This can be done either by quantifying *correlation* of user preference to individual input features of interest (e.g., user preference to presence of explosions in the image) or by *testing hypotheses* (e.g., “Is the user more likely to like an image if there is nudity?”).

Instead of analyzing the relation between individual input features X_i to output components Y_j , by applying *Canonical Correlation Analysis* it is also possible to find out which combination of features correlate with which combinations of output components.

Correlation coefficients

In its broadest sense, correlation refers to any statistical dependency between two random variables. More specifically, there exist several ways of calculating *correlation coefficients*, each one of them designed to measure the strength of a particular kind of statistical dependency. The most common candidates are Pearson’s correlation coefficient [33], which measures linear dependence between two continuous random variables, and Spearman’s rank correlation coefficient [34], which measures how well the relationship between the two variables can be described by a monotonic function. Both of these

coefficients are fairly simple to interpret, however, it shall be noted that a Pearson or Spearman coefficient of 0 does *not* imply the absence of any statistical dependency between the variables. For example, for X uniformly distributed on $[-3, 3]$ the random variables X and X^2 have Pearson and Spearman correlation 0 but are far from independent. There exist other correlation measures, which are able to capture more complex statistical dependencies but are typically harder to interpret. These include distance correlation introduced in [35], which is 0 only if the tested variables are independent. Specific choices should be made based on the properties of the tested variables (distributions they follow) and the questions one is trying to answer with the analysis.

Statistical significance

Correlation coefficients mainly measure the degree of a certain statistical dependency, but one should also check reliability of the findings by testing whether the dependency is *statistically significant*. This can be done based on *hypothesis testing* for estimating how likely it is that the true correlation is 0 (in a two-sided test, or ≤ 0 or ≥ 0 in one-sided tests) and the observed correlation value is due to noise. Another option is to calculate *confidence intervals* for the coefficients, for which a variety of methods have been proposed (e.g., see [36] for Spearman correlation).

Note that, coming directly from the definition of statistical dependency, we can also estimate confidence intervals for both the expected value of Y_j and the expected value of Y_j given a particular value x of X_i . If these confidence intervals do not overlap, this means that there is a significant difference between $E(Y_j)$ and $E(Y_j | X_i = x)$, i.e., X_i attaining value x significantly affects the expected value of the output Y_j . It shall be mentioned that overlapping confidence intervals do *not* imply that there is no significant difference [37].

Canonical Correlation Analysis (CCA)

Another set of statistical methods for analyzing the relation between two sets of variables (such as input and output variables in our case) is constituted by *Canonical Correlation Analysis*, or short CCA. The original CCA approach [38, 39] aims at finding linear relations between a matrix of input observations and a matrix of output observations. That is, if we are given a matrix M_X with m input features of n items and a matrix M_Y with o output values for the same n items as columns, the first objective is to find

two vectors z_X^1 and z_Y^1 that map M_X and M_Y on the n -dimensional unit ball such that the cosine similarity between $M_X z_X^1$ and $M_Y z_Y^1$ is maximized (i.e., the transformations of M_X and M_Y point in a similar direction). Iteratively, further vectors z_X^i and z_Y^i are calculated under the additional constraint that each z_X^i (and z_Y^i resp.) must be orthogonal to all previous vectors z_X^j (z_Y^j resp.), for all $j = 1, \dots, i - 1$ (see [40]).

Different from computing correlation coefficients between individual features, CCA belongs to multivariate statistics, and returns correlations between combinations of features. For example, in case of our toy example from Table 2, CCA gives us a result of the form

$$z_X^1 = (0.43, 0.85, -0.30)^T, \quad z_Y^1 = (-1),$$

indicating that a linear combination of 0.43 nudity, 0.85 humor and -0.3 explosions correlate maximally *negatively* with the user’s preference. Statistical significance of CCA results can be checked by applying Barlett’s sequential test procedure [41].

It is worth mentioning that CCA falls under dimensionality reduction techniques as well [42], a set of techniques which we will discuss below in Section 4.4. Furthermore, various modifications of CCA have been suggested. These include kernel-based [43] and neural network based methods [44, 45] for finding non-linear relations, as well as techniques that aim at improving interpretability of the discovered relations by enforcing sparsity on non-zero coefficients [46, 47, 48, 49]. For further details we refer to the comprehensive and recent tutorial on CCA by Uurtio et al. [40].

Remark on causality

Intuitively, we might wish to understand which features of the input *cause* a certain response. For example, an analysis of user preferences might ultimately aim at helping to design new contents by pointing at specific features that are linked to positive user reactions and thus are suggested to be incorporated. However, all methods we discussed try to figure out statistical dependence (correlation), which does not imply causation. In fact, causal assumptions can generally only be verified if experimental control is exerted [50]. In the general case described in this paper, the possibility for collecting additional data while manipulating parts of the input cannot be guaranteed. It shall be mentioned that for the case of analyzing given AI approaches, this possibility is likely to be given and there are some recent attempts in

computer science to address causality (e.g., [51, 52, 53, 54, 55]). We believe that this direction should be further explored for interpretation analysis in future work, and refer the interested reader also to the paper of Pearl [50] for a solid overview on causal inference in statistics.

Usage

Individual correlation coefficients are simple to understand, the methods for computing them are transparent and concrete statements about reliability can be made. Overall, correlation coefficients provide a robust way of quantifying the role of individual features as long as the feature space is not too high dimensional. On the downside, results crucially depend on selecting the right input and output features for analysis, which can be very challenging to do. This problem is less severe in CCA, which can pick up more complex dependencies. However, in comparison to correlation coefficients, canonical correlations tend to be harder to make sense of. An important advantage of statistical methods is that they allow for significance testing, which is necessary if specific claims in the form of hypotheses are to be tested rigorously.

4.2. Pattern mining

The general goal of pattern mining is to find characteristic patterns in the data. What exactly constitutes a pattern varies, but they often take on the the forms of association rules, emerging patterns or visual patches, as will be described in the following.

Association rule mining

Association rule mining has a long tradition in pattern mining [56, 57]. In particular, it is often used for web personalization where it is applied to usage data [58, 59, 60]. In its original form [57] it can be used to process a list of binary vectors and find implications of the form “if an image contains nudity and humor, then in 50% of cases the image also contains explosions” (using hypothetical data from Table 2).

Let $T = \{b_1, \dots, b_n\}$ be a multi-set of n transactions over k items represented as binary vectors with $b_i \in \mathbb{B}^k$, $n, k \in \mathbb{N}$. An association rule can formally be defined as implication of the form $X \Rightarrow j$, where $X \subseteq \{0, \dots, k\}$ is a set of indices called the antecedent of the rule, and $j \in \{0, \dots, k\} \setminus X$ is a single index (not included in X) called the consequent of the rule. The *support* of a set of indices X can then be defined as the relative amount of

transactions containing all items in X , and the *confidence* of a rule $X \Rightarrow j$ as the relative support of the rule’s antecedent and consequent over the support of its antecedent (see [57]):

$$\text{supp}(X) := \frac{|\{b_i \in T \mid b_{i,j} = 1, \forall j \in X\}|}{|T|} \quad (2)$$

$$\text{conf}(X \Rightarrow j) := \frac{\text{supp}(X \cup \{j\})}{\text{supp}(X)} \quad (3)$$

Another important measure *lift* [61], describes the ratio of the observed support for a rule to the support that would be expected if antecedent and consequent were independent. Confidence, support, other measures such as lift, and given potential constraints (e.g., only considering rules with specific j), can all serve as criteria for filtering possible rules. Association rules are often computed based on the apriori [62] or frequent pattern tree [63] algorithms (see e.g., the survey [64]).

For interpretation analysis, we are interested in so-called classification rules, i.e., rules that have a subset of the input as antecedent and a subset of the output as consequent [62]. So in our hypothetical example (Table 2), we would try to find rules of the form “if an image contains explosions, then the user likes it in 2/3 of cases.” Such a way of modeling is for instance adopted in [65], where association rule mining is used for finding class-discriminative features in images. In their approach, a binary class membership entry is appended to all vectors and only rules with this particular index as consequent are considered.

Emerging pattern mining

The problem of emerging pattern mining was introduced in [66], originally for capturing trends in time-stamped databases. It is similar to association rule mining, but uses the notion *growth rate* to measure how support for a pattern (set of indices) differs between sets. So broadly speaking, the goal of emerging pattern mining is to find differences in patterns across multiple sets. Soon after the task was introduced, it has been used for classification purposes [67, 68], where emerging patterns are meant to capture characteristic differences between classes. To this end, input samples are partitioned based on the associated output values and found patterns used to discriminate between the resulting partitions. It is in this sense that this approach

can directly be used for interpretation analysis. Coming back to our toy example of Table 2, following an emerging pattern mining approach we would ask, which are the combinations of nudity, humor and explosions that are comparatively more frequent in images the user likes/dislikes. Note that the survey of Novak et al. [69] puts emerging pattern mining under the umbrella term supervised descriptive rule discovery, together with contrast set mining and subgroup mining. Another useful resource is the recent survey of [70].

Visual pattern mining

There are several image-specific approaches worth mentioning. In [71], Rematas et al. use standard data mining terminology to formulate the problem of finding characteristic visual patches from a given image collection, which they also put into a graph for navigation through the image collection. The publications [72, 73] use association rule mining on mid-level CNN features, and call this combination mid-level deep pattern mining.

Note that sometimes the notion “parts” is used for referring to something comparable to visual patterns. For example, [74] describes how to automatically discover discriminative parts for the purpose of image classification. Visual pattern mining was also applied in [75], by using a bag-of-features representation (also known as bag-of-visual-words) [76] and selecting representative and discriminative local features based on Peng’s method for feature selection [77]. The recently proposed PatternNet [78] introduces a CNN that directly learns discriminative visual patterns. (As such, some of these approaches could as well be put into the model-based category described in the next section.)

Usage

Pattern mining approaches are conceptually similar to the statistical methods discussed above, as they discover relations between input and output features. The crucial difference is that in pattern mining approaches, these relations are described in different formats, which are designed to be intuitively understandable and can take the form of rules, discriminative patterns or characteristic visual patches. However, understanding can be hard for more complex patterns (e.g., very long rules) and while pattern mining techniques still include measures for reliability of the findings, there might be a high risk of ending up with many false alarms, since the space of possible patterns can be huge [79]. Also note that many pattern mining techniques operate on binary data, so it might become necessary to first convert the

data. In the above-mentioned paper [65], this is done for example by choosing a bag-of-features image representation. An example of an adaptation of pattern mining to textual data can be found in [80].

4.3. Model-based approaches

Even though the perspective of interest is considered to be a black-box in this paper, it is still possible to build another model to approximate the interpretation function based on the given input-output pairs. Successively, this trained model can be analyzed in the hope to reveal information about data dependencies that the original black-box might also rely on. In the example of our toy data (Table 2), we would first train a computational model to predict like/dislike from the input features nudity, humor and explosions, and successively analyze the trained model for dependencies between both parts.

We will discuss four kinds of model-based approaches which each focus on different aspects of analysis: First, *heatmapping* techniques aim at visually explaining decisions for individual items (e.g., “Which image features are likely to make the user like an individual image?”). Second, *prototype* approaches compute characteristic inputs for the different output classes (e.g., “How does a typical image look like which the user favors?”). Third, globally or partially *understandable models* can be used to approximate the perspective in order to obtain a more holistic understanding of how interpretation works. Finally, in *ablation studies*, the role of individual input features or model components is analyzed by removing them.

We do not go into too much detail for heatmapping and prototype methods because there are other survey papers such as [20, 9] which give an excellent overview for most of these approaches (in a non-black-box set-up). Similarly, [7] contains a comprehensive treatment of globally understandable models. Partially understandable models and ablation studies are less frequently mentioned in the context of explainability methods in machine learning research.

Heatmapping

In the context of analyzing machine learning models, a *heatmap* refers to an explanation of the model’s decision for a particular sample in terms of the input, indicating visually which parts of the input are relevant (positively or negatively) for the decision. For an example of a heatmap, see Figure 2. Heatmapping techniques can broadly be classified as methods for computing



Figure 2: Example of a heatmap computed from an inception-v3 [81] network which was trained for image classification on the imagenet dataset [82]. The top predicted class for this image was “malinois” (particular dog breed). The heatmap shows absolute values of input gradients, which serve as visual explanation of the classification result for this particular image. Smoothing and logscale have been applied to the gradients for illustration purposes. Image “Hey Big Dog!” by Alan Levine (<https://www.flickr.com/photos/cogdog/41916073004>, public domain).

saliency maps and *relevance methods*.

A common method is to calculate *saliency maps* [83, 84, 85] based on sensitivity analysis [86, 87, 88], i.e., the gradients on the model’s input are used for estimating how sensitive the model is to changes in the individual input components. A related approach is prediction difference analysis [89], which is still based on sensitivity analysis but uses local regularization in order to obtain visualizations that are easier to interpret. Saliency maps are simple to calculate for neural networks by means of backpropagation [90], but on the downside, resulting heatmaps have been shown to be unreliable in certain cases [91]. Also, sensitivities to input components is typically not exactly what we want to find out, because they only tell us how the input could be changed to make it belong more or less to a certain class instead of explaining which parts of the input actually make it belong to a class.

The latter can be achieved with *relevance methods* within the theoretical framework of Taylor decomposition [92]. In [93], Bach et al. adapt Taylor decomposition to neural networks and introduce layer-wise relevance propagation (LRP), which makes use of the network’s architecture to propagate

relevance backwards through the network for obtaining a heatmap. The backward propagation rule they derive takes two hyperparameters and for one particular combination, simplifies into a rule that is interpretable as deep Taylor decomposition [94]. Other backprop techniques have been proposed for computing heatmaps for neural networks, including Deconvolution [95], Guided Backprop [96], Class Activation Mapping [97], PatternAttribution [98] and PatternLRP [99].

Prototypes

Another way of visualizing what the model has learned is to calculate inputs that serve as prototypes for the individual classes. For example, a neural network trained on the MNIST dataset to recognize the digits 0-9 can be used to obtain a “typical” image for the digit 5. Prototypes can be calculated within the analysis framework of *activation maximization* [100, 101]. Essentially, finding prototypes amounts to solving the optimization problem of finding an input that maximizes a certain component of the output (e.g., an image that is interpreted by the model as being maximally dog-like). Without any additional restrictions, the resulting prototypes tend to be unnatural [85], which is why various regularization methods have been proposed [12, 102, 103]. For neural networks in particular, there are numerous efforts on visualizing what particular neurons or neuron layers have learned (e.g., [104, 95, 105]) which can also be seen as prototype approaches. An interesting non-prototype (but still related) approach is the one of [106, 107], where hidden unit activations are related to a binary segmentation task of the input for a given list of semantic concepts, in order to analyze semantics of individual hidden units.

Globally understandable models

Depending on the complexity of the data, it is possible to train a model that approximates the whole interpretation function in an understandable way. The most common candidates for such globally understandable models are linear models, decision trees and rules [7].

Linear models assign a weight to each feature, which provides a direct measure of the feature’s importance in terms of sign and magnitude. Especially in the social sciences it is common practice to use analysis of variance (ANOVA) [108] for analyzing experimental data. ANOVA is considered to be a special case of linear regression [109].

Decision trees are tree-like graphs, where internal nodes represent tests on input features and the leaf nodes represent a certain output. Decision trees have been used extensively since the early days of machine learning (see e.g., [110] and [111]). Similar to decision trees, decision sets [112] or decision lists [113] can be compiled from data. Note that conceptually many of these approaches are very closely related to association rule mining (which we discussed in Section 4.2). In fact, decision trees can be converted to sets of decision rules [114].

An interesting option that does not fall into any of these standard categories for understandable models are hypothesis-based models, which are common in the field of computational psychiatry [115]. There, conflicting hypotheses are implemented as computational models and fit on the given data to find which of the models (and therefore hypothesis) is better suited for explaining human information processing.

Partially understandable models

If the data is too complex for globally understandable models to fit properly, partially understandable models can be an appropriate compromise between understandability and prediction power. We discuss two ways of achieving this compromise: One is by breaking the problem down into more accessible steps in *pipeline approaches*, the other is to incorporate specific *structural components* into architectures that can be understood intuitively (e.g., explicit attention mechanism).

In pipeline approaches, specific mid-level features can be used to simplify understanding of the model’s output. For example, [116, 117] take the detour of recognizing adjective-noun combinations in images for the task of visual sentiment detection, and [118] propose a list of visual concepts to be used as intermediate features for classifying multimodal tweets of presumably gang-associated youth. Explicit attention mechanisms were mentioned above as one way to include understandable components into architectures. Such attention mechanisms are frequently used in machine translation [119], are a key component of memory networks [120, 121], and have been used for tasks such as image captioning as well [122]. A related approach is that of [123], which explains how to modify CNN architectures such that learned filters are more semantically meaningful and understandable.

Ablation studies

The principle of ablation studies is to gain understanding of the role of a system’s components by analyzing how the overall system changes if the component is removed. Historically, in neuroscience many early insights about functionality of individual brain regions were obtained by examining changes resulting from brain damage in particular areas [124]. In computer science, ablation studies have been adopted for quantifying the importance of model components [125], which for example can be used for model verification or reduction.

For interpretation analysis, ablation can be a useful tool when applied at the input level to address two points: First, which parts of the input are necessary for approximating the perspective of interest? If prediction performance drops drastically after removing a certain feature from the input, the feature was important for learning. This principle is frequently made use of in NLP for analyzing the role of features for prediction (e.g., for identifying hate speech [126]). Second, when having trained a model for perspective approximation, one might want to verify that the model does not use any parts of the input which it should not use (e.g., because they might be known not to be used by the original interpretation function). For example, [127] uses an ablation study where they mask the foreground to confirm that the classifier does not cheat by predicting from background properties.

Usage

In principle, model-based approaches can be used to learn complex dependencies, and heatmapping can explain decisions in individual cases, even when training models directly on pixel data [94] or word sequences [128]. Heatmapping has been applied together with several other features too, such as bag of visual words [93] and fisher vectors [129]. Such model-based explanations were found to be useful in many publications (e.g., [15, 14, 89]). Zhou et al. [97] also show how a network can learn to localize objects with decent performance without any bounding box labels.

Still, in general it is not clear which properties of the original perspective carry over to the trained model when fitting it on a given list of inputs and outputs, and to the best of our knowledge, there is no extensive study analyzing the transfer of various functional properties. Indeed, publications dealing with adversarial noise (e.g., [130, 131, 132]) show how convolutional neural networks are typically sensitive to things which humans are not [133, 134, 11], despite being trained on large amounts of humanly annotated data

and convolutional neural networks originally being inspired by human vision [135]. This gives reason for caution when making claims about the original function based on analyzing its approximation, especially also for complex approximation methods such as deep neural networks. If the models are simpler and do not have the capacity for picking up on any complex noise, some of these issues can be ruled out and the approach becomes closer to statistical testing. Other partial remedies are to rely on pipeline approaches, where individual steps can be verified separately, or make use of ablation studies to rule out certain unwanted properties. Still, one should not confuse the trained model with the original perspective of interest, and be aware that there often is a remaining risk that findings are unreliable or misleading.

4.4. Visualization techniques

In the following, we describe visualization techniques in a very broad sense as methods to obtain a condensed representation of some given data. This representation can take various forms: In *dimensionality reduction*, the data is transformed into a lower-dimensional space such that it can be plotted. Other methods stay closer to the original type of data and rather reduce the amount of information in different ways. These include extraction of *examples*, reducing the amount of information by topic modeling, or automatic *summarization*.

Dimensionality reduction

Dimensionality reduction can be useful for visualizing almost any kind of data by reducing the data dimension, such that it can then be plotted and manually inspected. There are many different kinds of dimensionality reduction and several surveys have been made on the topic [136, 137, 42]. Here, we outline a few popular cases that are especially relevant for interpretation analysis. Linear dimensionality reduction refers to methods that linearly transform the original input space, i.e., they describe how to find a matrix that is multiplied to all inputs for projecting them into a smaller space (see [42]). Popular methods that fall into this category are Principal Component Analysis (PCA) [138], Linear Discriminant Analysis (LDA) (e.g., [139]), and Canonical Correlation Analysis (CCA) [39], which all compute orthogonal matrices for the transformation. LDA uses associated class labels and transforms the input space such that after transformation the separation between the classes is maximized. This is closely related to linear regression, which can be seen as another linear dimensionality reduction technique that

does not use an orthogonality constraint. An interesting property of PCA is that after transformation the components are linearly uncorrelated or, in other words, the data is factorized into independent components. Other popular factorization methods include Factor Analysis (FA) [140], which is widely used in psychology [141], for example to become aware of patterns in questionnaire items [142].

Linear dimensionality reduction with orthogonal matrices can be especially helpful for getting a rough idea of the data's structure, since they do not exaggerate relations between data points (see [42]). Projections of non-linear transformation techniques can be harder to interpret since geometric properties like distances in the original space are generally not preserved. Still, such techniques can be useful for looking at specific properties of the data, and there are a few non-linear transformation techniques that deserve mentioning: t-SNE [143] is a probabilistic method that embeds samples into a low-dimensional space such that similar samples are likely to be embedded to nearby points and dissimilar object to distant points. Another non-linear reduction technique is to train an autoencoder [144] to compress the original data into a smaller latent encoding. The benefit of autoencoders is that they can be combined with additional loss functions for enforcing other properties on these encodings, such as following a certain distribution [145] or using specific positions to encode certain semantic properties [146].

Example-based approaches

The idea behind example-based approaches is that even for large collections, looking at characteristic examples can be useful to automatically form a holistic understanding of the collection. The crux herein is to select the right examples (and know how many are necessary), for which various approaches exist.

A simple and yet useful method is to randomly select a few samples for manual inspection. This cannot be expected to lead to a full understanding of the sample collection but helps to form an initial feeling for the data. One issue is the possibility that by chance odd samples are drawn, which are included in the data, but exhibit certain unexpected properties. Obtaining such abnormal examples can also be done on purpose, which relates to a common task called anomaly detection (see e.g., [147]). Anomalies can for example help to become aware of problems with the data (e.g., broken entries), but can also be of particular relevance when working with methods that are sensitive to statistical outliers (e.g., linear regression).

There are other ways how samples can “stand out” and hence be interesting to look at. For example, the sample which is closest to the average over all samples can be seen as most representative of the whole set, or, if there are different output scores, it is sensible to look at a few samples with different scores. Other sophisticated methods exist to obtain representative and diverse examples for visualizing sample collections. For image collections, summarization is most commonly done by selecting representative examples. For example, in [148] the selection of representative images is formulated as optimization problem and mixtures of submodular functions are learned for scoring selections. In [149] the authors extract SIFT features and use a modification of RANSAC [150] plus Affinity Propagation clustering [151] for finding representative images. If there is accompanying textual or social information for the images, other approaches exist (e.g., see [152, 153, 154]).

Text summarization

For textual data, visualization and summarization techniques have been extensively surveyed [155, 156, 157, 158, 159], and it is commonly distinguished between extractive techniques and abstractive techniques. Extractive summarization techniques aim at compiling a list of sentences (examples) that summarize the collection. Abstractive summarization techniques include the extraction of topic words, frequency-driven approaches such as tf-idf, and automatic summarization. It is important to note that in our context, we generally not only want to summarize all the given inputs, but summarize in a way that reveals differences between between inputs associated with different outputs. Specific works on discriminative text summarization include [160], which explains how to select discriminative sentences for summarizing differences between text collections, and [161], which aims at visualizing differences between text corpora based on discriminative words or by analyzing an SVM that was trained to detect the source of the text.

Usage

Note that ultimately, in interpretation analysis we are not interested in merely visualizing the collection of inputs or outputs, but to do so in a way that shows relations between input and output values. There are three main ways how this can be achieved: 1) If we want to apply dimensionality reduction to the input, the associated values can directly be incorporated into the visualization, e.g., by using colors to indicate different associated output values. 2) For applying dimensionality reduction to the output, if we have

(short) text data or images as input data it is possible to show the original inputs at the locations of their corresponding output embeddings. 3) Finally, for example-based approaches and text summarization, input samples can be partitioned based on associated values for separate visualization and successive comparison of results.

Visualization techniques can be very beneficial for an intuitive understanding of perspective, and can serve as useful starting point for getting ideas about which features to explore or which types of hypotheses to test with quantitative methods. On the downside, it is hard to draw any concrete conclusions from visualizations alone.

5. Comparing multiple perspectives

Understanding differences between various perspectives has use-cases in a variety of scenarios. For example, one might be interested in the difference between two given machine learning classifiers, understanding how distinct annotators label data differently, comparing a classifier’s perspective to the ground truth human perspective, or analyzing in which ways data from different domains relates to interpretation-related discrepancies. Even if one is not directly interested in such a comparative study and the ultimate interest is only in understanding one given perspective, it is sensible to compare against a baseline perspective for making results easier to interpret. For example, it seems that the user in our toy example (Table 2) slightly prefers explosions in images, but perhaps everyone has such a preference? Maybe what’s really special about this user’s interpretation is that nudity or humor do not seem to affect her preference in clear way?

So, assume we are given several lists of inputs and their corresponding outputs, each list being associated to one perspective, and we want to characterize in which ways the underlying interpretation functions are different. For example, in our image preference scenario, we can imagine to be given similar tables from other users and want to see how their preferences differ. To this end, individual perspectives can be analyzed separately and then compared, one can merge the perspectives into a single one and then analyze, or combine all perspectives in a single model. We discuss all of these possibilities for comparison below. An overview can be found in Figure 3.

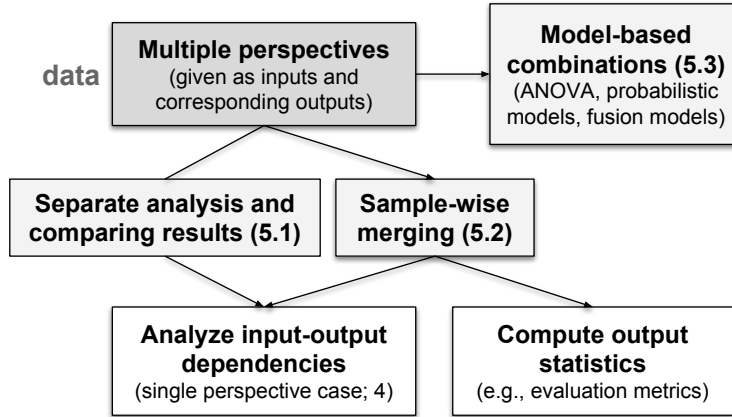


Figure 3: Approaches for comparing multiple ways of interpretation. We can distinguish between three possibilities, out of which two mainly reduce the comparison problem to the analysis of a single perspective.

5.1. Comparing input-output dependencies

Most recent papers that aim at explaining differences of machine learning models first analyze input-output dependencies by using model-based approaches mentioned above (Section 4.3), and then compare the results, typically by displaying them side by side (see [9]). Such an approach of separately analyzing individual perspectives followed by comparison can be seen as direct attempt to answer the question “How do relations between inputs and outputs differ across the given perspectives?”

Conceptually this offers a simple way to compare, but can suffer from several issues: Findings for the different ways of interpretation might be very similar and differences not at all apparent. For instance, if for user A we find a single rule “nudity and explosions lead to like in 70% of cases” while for user B we obtain “explosions lead to like in 60% of cases”, then what exactly is the difference between their ways of interpretation? Also, if there are many interpretation functions, but only little data for each, analyzing individual perspectives might be unfeasible or not give any significant results. For example, in the context of recommender systems we might only have 5 items rated per user, which is insufficient for complex statistical analyses or using model-based approaches (on a single user).

Despite these potential shortcomings, there are cases where it makes per-

fect sense to analyze perspectives separately and then compare. Most importantly, often there is an interest in understanding individual ways of interpretation as well. In such cases, individual perspectives would typically be analyzed anyway, so comparing results would only cause little computational overhead and thus provides a reasonable starting point. When facing any of the above-mentioned issues, one can still follow up with sample-wise or model-based combinations, which we discuss in the remainder of this section.

5.2. Sample-wise combinations

We can phrase the slightly different question “How do inputs relate to differences in the outputs?” Let us first assume we have function values from two different interpretation functions f_{b_1}, f_{b_2} on the same set of input samples i_1, \dots, i_n . We can easily define a new perspective f that is described by the same input samples and their associated outputs $d(f_{b_1}(i_1), f_{b_2}(i_1)), \dots, d(f_{b_1}(i_n), f_{b_2}(i_n))$, where d is any real-valued vector function that calculates a difference or distance between two values, e.g., $d(y_1, y_2) = |y_1 - y_2|$. Thereby, the function d should be chosen depending on the overall goal: If one is only interested in finding out explanations for when there is disagreement between the two perspectives, one might want to choose a binary indicator of equality, or the absolute value of the difference between both outputs. If the goal is to also understand the direction of disagreement, the mere difference without absolute value is more suitable. For example, if we are given two computer models A and B for sentence-level sarcasm detection, we might ask which features of the sentence are related to any disagreement between A and B (binary case), but we can also analyze which features make model A but not B vote for sarcasm.

Irrespective of the choice of the merging function d , this resulting perspective f can be analyzed as in the single perspective case. This is a straightforward way to directly analyze differences between ways of interpretation, and checking statistical significance works in the same way as for a single perspective. For such a merged perspective, output statistics can be computed too, for example in order to evaluate a learned perspective f_{b_1} against a target perspective f_{b_2} . The case of comparing more than two interpretation functions can be handled analogously.

Remark – performance measures

Many performance measures can be seen as sample-wise combination approaches, where the perspective of the classifier is compared to the perspec-

tive given by the ground truth labels. Typically these measures combine the perspectives in fairly simple ways. For example, accuracy would use a binary equality indicator as d and average over all outputs of the merged perspective, precision would use the same d but average only over a certain part of the outputs (the ones where the interpretation of the classifier was positive).

5.3. Model-based combinations

Another possibility is to combine several perspectives in a single model. This relates to the question “How does the bearer influence interpretation?” Models for model-based combination of perspectives can take various forms, three of which we are going to discuss in this section.

ANOVA

A simple case would be the use of ANOVA, with interpretation output as dependent variable and both input features and identifier of the interpretation function (or features that group them, such as demographic information) as independent variables. ANOVA would then tell us whether there is a significant difference among average output values across the perspectives.

Probabilistic models

Even though less common, there are more complex possibilities for combining perspectives in probabilistic models. Typically, the main goal of such probabilistic models is not to analyze ways of interpretation, but to learn how to combine multiple perspectives for a given prediction task. Still, characteristic information about the involved perspectives can be picked up by such models. An example of such a probabilistic model for combining human perspectives is the Dawid-Skene model [162], which unites observations from different sources while estimating the observers’ errors. Further examples will be given in the application section (Section 6.3).

Fusion models

For AI approaches, ensemble methods are frequently used for increasing predictive performance [163]. These methods often include a scoring mechanism or allow for similar ways of obtaining an estimation of the usefulness of the individual models involved, which can be seen as discriminative characterization.

Another approach to fusion is taken in end-to-end fusion models, where a single model (usually a neural network) is trained to predict the interpretation result given the input and information about the bearer. This

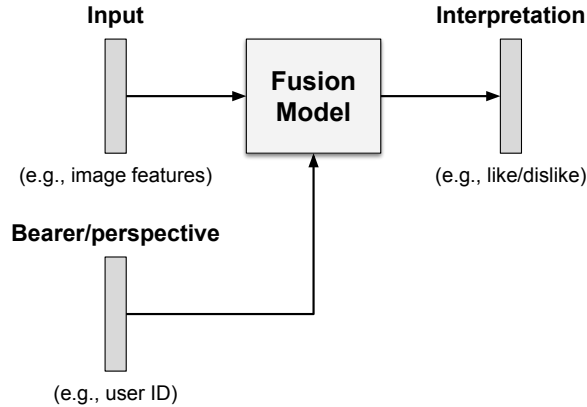


Figure 4: Illustration of end-to-end fusion models for comparing perspectives. Such models are first trained to predict interpretation results from input and information about the bearer. The trained model is then used for analysis.

end-to-end approach is illustrated in Figure 4. This way of combining information is relatively common for prediction but has rarely been used for the purpose of analysis. Possibilities for analysis include heatmapping and prototype techniques (see Section 4.3). Additionally, end-to-end fusion offers extra opportunities such as learning vector representations for the individual perspectives which can be used for clustering for example. However, extra care should be taken when interpreting findings based on such complex models.

6. Applications

Tools for interpretation analysis can be utilized in a variety of scenarios. In the following, we outline some of the cornerstones.

6.1. Mining subjective information

Prominent examples of applications that aim at mining subjective information from text data are sentiment analysis and opinion mining [164, 165]. The main task of sentiment analysis is to decide whether a given text expresses a positive, a negative, or a neutral opinion, which can for example be useful for evaluating customer reviews. In its original form, sentiment analysis is about learning a way of interpretation, but does not necessarily involve

any claims about characteristics of the same. However, it is very common to not simply detect overall sentiment, but to do so based on aspects. The resulting detection pipeline then has aspect information as extra component, and tries to explain the overall sentiment in terms of mentioned aspects and the orientation expressed towards these. For understanding persisting differences in interpretation, contrastive opinion mining has been proposed by Fang et al. [166] and, later, perspective detection by Vilares and He [167]. The Latent Argument Model in [167] is a rather complex case of discriminative text summarization based on topic modeling, and is paired in the paper with selection of characteristic sentences. Note that sentiment analysis was extended to the visual modality as well. Somewhat similar to aspect-based sentiment detection, Borth et al. proposed a visual sentiment ontology [116] consisting of adjective-noun combinations (e.g., “scary dog”, “cute baby”) that are visually detectable and can be used for explaining the overall sentiment of an image.

Quite a different approach is taken in [168], which analyzes how hotel preferences change over time by applying emerging pattern mining on hotel features mentioned in online reviews.

6.2. Model analysis

Several papers have explored the possibility to use decision trees for explaining more complex machine learning models, including neural networks [169, 170, 171] and tree ensembles [172, 173]. Furthermore, much recent work was done on analyzing deep learning models and explaining decisions based on heatmapping (e.g., [89, 94, 99, 98]). These are all direct cases of model-based interpretation analysis (usually not operating under the same black-box assumption though). Visualization techniques have been used as well for examining learned representations of neural networks. For example, [174] use t-SNE on phrase embeddings (which can be seen as output of the model’s interpretation function) to analyze how semantically meaningful the learned embeddings are.

Note that computation of many performance metrics can be seen as special case of interpretation analysis, where the output of a classifier is compared to a ground truth human interpretation by merging both perspectives in a sample-wise manner and then aggregating over the outputs of this combined perspective.

6.3. Annotation

Computer vision in particular depends on big amounts of manually labelled data for training models, which is often achieved via crowdsourcing [175]. In crowdsourcing, it is common to collect several annotations for each item, and many probabilistic models for merging annotator votes have been proposed (e.g., [176, 177, 178, 17, 18]). Often, these simultaneously estimate annotator reliability, but only a few approaches consider item difficulty and thereby relate disagreements to the input. Notable exceptions are [17] and its extension [18], which describe such a probabilistic framework and apply their framework to merge fine-grained bird image annotations. Less work has been done on investigating where annotator disagreements come from. One of the few examples in this direction is [179], which analyzes correlations between textual and visual item features and annotator disagreement in case of labeling multimodal tweets as *aggression*, *loss* and *substance use*. For crowdsourcing, Eickhoff [180] outlines several quality issues and performs dedicated experiments for analyzing cognitive biases of annotators. The paper also shows how such biases can propagate into model evaluation and hence have detrimental consequences, which gives reason for further investigation into a more fine-grained interpretation analysis for annotation.

6.4. Data understanding and expertise

In many scientific undertakings the goal is to understand the relation between two quantities based on some given data. Interpretation analysis tools have been applied to make sense of various kinds of scientific data. Early examples include the application of CCA to describe the relation between wheat to flour characteristics [181] or to analyze how housing quality interacts with mental issues [182]. As another example, association rule mining has been used for making sense of gene expression data [183] and medical data [184]. Emerging pattern mining for finding differences between toxic vs non-toxic chemicals [185]. Visual pattern mining for histology image collections is done in [75] for identifying local features that can be used to discriminate between tissue types. The same paper also estimates posterior probabilities for relating local features to individual tissue types for interpretation. Numerous attempts at data explanation have also been made by fitting various models on the given data and then analyzing the trained models for insights. In [15], a deep tensor neural network model with heatmapping was applied to examine the link from molecular structure to electronic properties. And [14] reported LRP-based explanations for classifying EEG data

with a neural network to be highly plausible. Essentially, such cases can be seen as figuring out some “natural” way of interpretation that is intrinsic to the given data. In the special case when the output quantity is given in the form of labels from human experts, analyzing the data amounts to explaining their expert view, or in other words, to characterize an expert’s way of interpretation. Note, however, that for data understanding our black-box assumption (see Section 2.2) is generally satisfied, so care has to be taken when interpreting the trained model.

6.5. Understanding human interpretation

Mechanisms and properties of human interpretation are of fundamental interest in several fields, including cognitive science, neuroscience, phenomenology, linguistics, psychology and psychiatry. Traditionally, these fields often conduct designated controlled experiments for data collection, or use qualitative analysis when relying on given observational data. Still, there are some approaches that are more in between the fields mentioned above and computer science. These include recent works on computational psychiatry [186, 187, 115] which turn hypothesis about human functioning into simple computational models that can be evaluated on experimental or observational data. For example, [115] explains how to use a hierarchical generative model for exploring potential relations between over-attention to low-level stimuli and schizophrenia. Another model-based approach is taken in [188] for studying language acquisition by feeding language data into a model based on hidden Markov model. Their trained model is then evaluated by comparing the model’s word generalization abilities against the ones of children [189], and can be useful for generating predictions about language development.

7. Conclusion

In this paper, we proposed a theoretical framework in which we formally defined interpretation, perspective and the task of interpretation analysis. In our framework, interpretation analysis can be understood as characterizing functions and describes relations between inputs and corresponding outputs. We showed how analyzing a single way of interpretation can be approached under the use of statistical methods, pattern mining techniques, model-based approaches and visualization techniques. We discussed how comparing several ways of interpretation can often be reduced to the single perspective case, and alternatively be handled by uniting perspectives in a designated

model for analysis. Finally, we have seen applications from several areas, including opinion mining, annotation and analysis of machine learning models, which can be connected by their relations to interpretation analysis.

During our survey of approaches, we identified several points that we think deserve more attention in the future. In particular, proper evaluation of interpretation analysis methods is still largely an open issue. This holds true especially for more complex model-based approaches under our black-box assumption (generally satisfied when using them for data understanding) and visualization techniques. Further, there are many qualitative methods that are relevant to interpretation analysis which we hope can further inspire computational methods in the future. Similarly, though we have already drawn many connections between literature from the fields of behavioural sciences, psychology and computer science in this paper, we hope to see more work in the fruitful intersection of these fields in the future. Last but not least, we see an ever increasing need for ethical discussions: Many application areas of interpretation analysis ethically concern user privacy. Similar techniques to the ones described have in the recent past already been used for ethically very questionable goals under the term microtargeting (e.g., to influence the outcome of elections [23]). Our hope is that the scientific community will in the future focus on using the same techniques for ethically less questionable goals, for example to increase transparency and explainability of AI systems and maybe even to help us become aware of our own detrimental biases.

Acknowledgments

This work was supported by the BMBF project DeFuseNN (grant number 01IW17002) and the NVIDIA AI Lab (NVAIL) program. Furthermore, the first author received financial support from the Center for Cognitive Science, Kaiserslautern, Germany.

References

- [1] M. V. Campos, A. M. L. Gutiérrez, The notion of point of view, in: *Temporal Points of View*, Springer, 2015, pp. 1–57 (2015).
- [2] J. Zimmermann, *Hermeneutics: A very short introduction*, OUP Oxford, 2015 (2015).

- [3] J. P. Forgas, K. D. Williams, S. M. Laham, W. Von Hippel, et al., *Social motivation: Conscious and unconscious processes*, Vol. 5, Cambridge University Press, 2005 (2005).
- [4] J. Poushter, et al., *Smartphone ownership and internet usage continues to climb in emerging economies*, Pew Research Center 22 (2016) 1–44 (2016).
- [5] EU Council, *EU Regulation 2016/679 General Data Protection Regulation (GDPR)*, Official Journal of the European Union 59 (2016) 1–88 (2016).
URL <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679>
- [6] B. Goodman, S. Flaxman, *EU regulations on algorithmic decision-making and a "right to explanation"*, in: *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, 2016, pp. 1–9 (2016).
arXiv:1606.08813.
URL <http://arxiv.org/abs/1606.08813>
- [7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, *A survey of methods for explaining black box models*, *ACM computing surveys (CSUR)* 51 (5) (2018) 93 (2018).
- [8] W. Samek, T. Wiegand, K.-R. Müller, *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*, *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services* 1 (1) (2018) 39–48 (2018).
URL <https://www.itu.int/en/journal/001/Pages/05.aspx>
- [9] G. Montavon, W. Samek, K.-R. Müller, *Methods for interpreting and understanding deep neural networks*, *Digital Signal Processing* 73 (2018) 1 – 15 (2018). doi:<https://doi.org/10.1016/j.dsp.2017.10.011>.
URL <http://www.sciencedirect.com/science/article/pii/S1051200417302385>
- [10] Z. C. Lipton, *The mythos of model interpretability*, *Queue* 16 (3) (2018) 30:31–30:57 (Jun. 2018). doi:10.1145/3236386.3241340.
URL <http://doi.acm.org/10.1145/3236386.3241340>

- [11] S. Palacio, J. Folz, J. Hees, F. Raue, D. Borth, A. Dengel, What do deep networks like to see?, in: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018 (June 2018).
- [12] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196 (2015).
- [13] A. Mahendran, A. Vedaldi, Visualizing deep convolutional neural networks using natural pre-images, *International Journal of Computer Vision* 120 (3) (2016) 233–255 (2016).
- [14] I. Sturm, S. Lapuschkin, W. Samek, K.-R. Müller, Interpretable deep neural networks for single-trial eeg classification, *Journal of neuroscience methods* 274 (2016) 141–145 (2016).
- [15] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, *Nature communications* 8 (2017) 13890 (2017).
- [16] B. Alipanahi, A. DeLong, M. T. Weirauch, B. J. Frey, Predicting the sequence specificities of dna-and rna-binding proteins by deep learning, *Nature biotechnology* 33 (8) (2015) 831 (2015).
- [17] S. Branson, G. Van Horn, P. Perona, Lean crowdsourcing: Combining humans and machines in an online system, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7474–7483 (2017).
- [18] G. Van Horn, S. Branson, S. Loarie, S. Belongie, P. Perona, Lean multiclass crowdsourcing, in: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018 (June 2018).
- [19] A. M. L. Gutiérrez, M. V. Campos, Subjective and objective aspects of points of view, in: *Temporal Points of View*, Springer, 2015, pp. 59–104 (2015).
- [20] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, *IEEE Transactions on Neural Networks and Learning Systems* 28 (11)

- (2017) 2660–2673 (2017). doi:10.1109/TNNLS.2016.259982.
 URL <http://dx.doi.org/10.1109/TNNLS.2016.259982>
- [21] S. Escalera, X. Baró, H. J. Escalante, I. Guyon, Chalearn looking at people: A review of events and resources, in: *Neural Networks (IJCNN)*, 2017 International Joint Conference on, IEEE, 2017, pp. 1594–1601 (2017).
- [22] S. C. Woolley, P. N. Howard, Automation, algorithms, and politics—political communication, computational propaganda, and autonomous agents introduction, *International Journal of Communication* 10 (2016) 9 (2016).
- [23] F. J. Zuiderveen Borgesius, J. Moller, S. Kruikemeier, R. Ó. Fathaigh, K. Irion, T. Dobber, B. Bodo, C. de Vreese, Online political microtargeting: Promises and threats for democracy, *Utrecht L. Rev.* 14 (2018) 82 (2018).
- [24] M. Zook, S. Barocas, K. Crawford, E. Keller, S. P. Gangadharan, A. Goodman, R. Hollander, B. A. Koenig, J. Metcalf, A. Narayanan, et al., Ten simple rules for responsible big data research, *PLoS computational biology* 13 (3) (2017) e1005399 (2017).
- [25] Wired, Predictive Policing: Using Machine Learning to Detect Patterns of Crime (2013).
 URL <https://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/>
- [26] L. McClendon, N. Meghanathan, Using machine learning algorithms to analyze crime data, *Machine Learning and Applications: An International Journal (MLAIJ)* 2 (1) (2015) 1–12 (2015).
- [27] MIT Technology Review, Police across the US are training crime-predicting AIs on falsified data (2019).
 URL <https://www.technologyreview.com/s/612957/predictive-policing-algorithms-ai-crime-dirty-data/>
- [28] D. Freelon, C. D. McIlwain, M. Clark, Beyond the hashtags: # ferguson, # blacklivesmatter, and the online struggle for offline justice, Center for Media & Social Impact, American University, Forthcoming (2016).

- [29] D. U. Patton, D.-W. Brunton, A. Dixon, R. J. Miller, P. Leonard, R. Hackman, Stop and frisk online: Theorizing everyday racism in digital policing in the use of social media for identification of criminal conduct and associations, *Social Media + Society* 3 (3) (2017) 2056305117733344 (2017). [arXiv:https://doi.org/10.1177/2056305117733344](https://doi.org/10.1177/2056305117733344), [doi:10.1177/2056305117733344](https://doi.org/10.1177/2056305117733344).
URL <https://doi.org/10.1177/2056305117733344>
- [30] D. J. Levitin, Experimental design in psychological research, in: *Foundations of cognitive psychology: Core readings*, MIT Press, 2002, pp. 115–130 (2002).
- [31] P. Domingos, A few useful things to know about machine learning, *Communications of the ACM* 55 (10) (2012) 78–87 (2012).
- [32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
- [33] K. Pearson, Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London* 58 (1895) 240–242 (1895).
- [34] A. D. Well, J. L. Myers, *Research design & statistical analysis*, Psychology Press, 2003 (2003).
- [35] G. J. Székely, M. L. Rizzo, N. K. Bakirov, et al., Measuring and testing dependence by correlation of distances, *The annals of statistics* 35 (6) (2007) 2769–2794 (2007).
- [36] J. Ruscio, Constructing confidence intervals for spearman's rank correlation with ordinal data: a simulation study comparing analytic and bootstrap methods, *Journal of Modern Applied Statistical Methods* 7 (2) (2008) 7 (2008).
- [37] A. Knezevic, Overlapping confidence intervals and statistical significance, *StatNews: Cornell University Statistical Consulting Unit* 73 (1) (2008).
- [38] H. Hotelling, The most predictable criterion., *Journal of educational Psychology* 26 (2) (1935) 139 (1935).

- [39] H. Hotelling, Relations between two sets of variates, *Biometrika* 28 (3/4) (1936) 321–377 (1936).
- [40] V. Uurtio, J. M. Monteiro, J. Kandola, J. Shawe-Taylor, D. Fernandez-Reyes, J. Rousu, A tutorial on canonical correlation methods, *ACM Computing Surveys (CSUR)* 50 (6) (2018) 95 (2018).
- [41] M. S. Bartlett, The statistical significance of canonical correlations, *Biometrika* 32 (1) (1941) 29–37 (1941).
- [42] J. P. Cunningham, Z. Ghahramani, Linear dimensionality reduction: Survey, insights, and generalizations, *The Journal of Machine Learning Research* 16 (1) (2015) 2859–2900 (2015).
- [43] P. L. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, *International Journal of Neural Systems* 10 (05) (2000) 365–377 (2000).
- [44] P. L. Lai, C. Fyfe, A neural implementation of canonical correlation analysis, *Neural Networks* 12 (10) (1999) 1391–1397 (1999).
- [45] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: *International conference on machine learning*, 2013, pp. 1247–1255 (2013).
- [46] E. Parkhomenko, D. Tritchler, J. Beyene, Genome-wide sparse canonical correlation of gene expression with genotypes, in: *BMC proceedings*, Vol. 1, BioMed Central, 2007, p. S119 (2007).
- [47] S. Waaijenborg, P. C. V. de Witt Hamer, A. H. Zwinderman, Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis, *Statistical applications in genetics and molecular biology* 7 (1) (2008).
- [48] D. M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics* 10 (3) (2009) 515–534 (2009).
- [49] D. R. Hardoon, J. Shawe-Taylor, Sparse canonical correlation analysis, *Machine Learning* 83 (3) (2011) 331–353 (2011).

- [50] J. Pearl, Causal inference in statistics: An overview, *Statist. Surv.* 3 (2009) 96–146 (2009). doi:10.1214/09-SS057.
URL <https://doi.org/10.1214/09-SS057>
- [51] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, B. Schölkopf, Nonlinear causal discovery with additive noise models, in: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21*, Curran Associates, Inc., 2009, pp. 689–696 (2009).
URL <http://papers.nips.cc/paper/3548-nonlinear-causal-discovery-with-additive-noise-models.pdf>
- [52] D. Lopez-Paz, K. Muandet, B. Schölkopf, I. Tolstikhin, Towards a learning theory of cause-effect inference, in: *International Conference on Machine Learning*, 2015, pp. 1452–1461 (2015).
- [53] K. Zhang, A. Hyvärinen, On the identifiability of the post-nonlinear causal model, in: *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, AUAI Press, 2009, pp. 647–655 (2009).
- [54] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, K. Bollen, Directlingam: A direct method for learning a linear non-gaussian structural equation model, *Journal of Machine Learning Research* 12 (Apr) (2011) 1225–1248 (2011).
- [55] O. Stegle, D. Janzing, K. Zhang, J. M. Mooij, B. Schölkopf, Probabilistic latent variable models for distinguishing between cause and effect, in: *Advances in Neural Information Processing Systems*, 2010, pp. 1687–1695 (2010).
- [56] G. Piateski, W. Frawley, *Knowledge discovery in databases*, MIT press, 1991 (1991).
- [57] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: *Acm sigmod record*, Vol. 22, ACM, 1993, pp. 207–216 (1993).
- [58] M. Eirinaki, M. Vazirgiannis, Web mining for web personalization, *ACM Transactions on Internet Technology (TOIT)* 3 (1) (2003) 1–27 (2003).

- [59] B. Mobasher, R. Cooley, J. Srivastava, Automatic personalization based on web usage mining, *Communications of the ACM* 43 (8) (2000) 142–151 (2000).
- [60] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Effective personalization based on association rule discovery from web usage data, in: *Proceedings of the 3rd international workshop on Web information and data management*, ACM, 2001, pp. 9–15 (2001).
- [61] S. Brin, R. Motwani, J. D. Ullman, S. Tsur, Dynamic itemset counting and implication rules for market basket data, *Acm Sigmod Record* 26 (2) (1997) 255–264 (1997).
- [62] R. Agrawal, R. Srikant, et al., Fast algorithms for mining association rules, in: *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215, 1994, pp. 487–499 (1994).
- [63] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, in: *ACM sigmod record*, Vol. 29, ACM, 2000, pp. 1–12 (2000).
- [64] Q. Zhao, S. S. Bhowmick, *Association rule mining: A survey*, Nanyang Technological University, Singapore (2003).
- [65] T. Quack, V. Ferrari, B. Leibe, L. Van Gool, Efficient mining of frequent and distinctive feature configurations, in: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, IEEE, 2007, pp. 1–8 (2007).
- [66] G. Dong, J. Li, Efficient mining of emerging patterns: Discovering trends and differences, in: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 1999, pp. 43–52 (1999).
- [67] G. Dong, X. Zhang, L. Wong, J. Li, Caep: Classification by aggregating emerging patterns, in: *International Conference on Discovery Science*, Springer, 1999, pp. 30–42 (1999).
- [68] J. Li, G. Dong, K. Ramamohanarao, Instance-based classification by emerging patterns, in: *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, 2000, pp. 191–200 (2000).

- [69] P. K. Novak, N. Lavrač, G. I. Webb, Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining, *Journal of Machine Learning Research* 10 (Feb) (2009) 377–403 (2009).
- [70] A. Garca-Vico, C. Carmona, D. Martn, M. Garca-Borroto, M. del Jesus, An overview of emerging pattern mining in supervised descriptive rule discovery: taxonomy, empirical study, trends, and prospects, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8 (1) (2018) e1231 (2018). arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1231>, doi:10.1002/widm.1231. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1231>
- [71] K. Rematas, B. Fernando, F. Dellaert, T. Tuytelaars, Dataset fingerprints: Exploring image collections through data mining, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4867–4875 (2015).
- [72] Y. Li, L. Liu, C. Shen, A. van den Hengel, Mid-level deep pattern mining, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 971–980 (2015).
- [73] Y. Li, L. Liu, C. Shen, A. v. d. Hengel, Mining mid-level visual patterns with deep cnn activations, *International Journal of Computer Vision* 121 (3) (2017) 344–364 (Feb 2017). doi:10.1007/s11263-016-0945-y. URL <https://doi.org/10.1007/s11263-016-0945-y>
- [74] S. N. Parizi, A. Vedaldi, A. Zisserman, P. Felzenszwalb, Automatic discovery and optimization of parts for image classification, arXiv preprint arXiv:1412.6598 (2014).
- [75] A. Cruz-Roa, J. C. Caicedo, F. A. González, Visual pattern mining in histology image collections using bag of features, *Artificial intelligence in medicine* 52 (2) (2011) 91–106 (2011).
- [76] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: *Workshop on statistical learning in computer vision, ECCV*, Vol. 1, Prague, 2004, pp. 1–2 (2004).

- [77] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Transactions on pattern analysis and machine intelligence* 27 (8) (2005) 1226–1238 (2005).
- [78] H. Li, J. G. Ellis, L. Zhang, S.-F. Chang, Patternnet: Visual pattern mining with deep neural network, in: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR '18*, ACM, New York, NY, USA, 2018, pp. 291–299 (2018). doi: 10.1145/3206025.3206039.
URL <http://doi.acm.org/10.1145/3206025.3206039>
- [79] G. I. Webb, Discovering significant patterns, *Machine learning* 68 (1) (2007) 1–33 (2007).
- [80] N. Zhong, Y. Li, S.-T. Wu, Effective pattern discovery for text mining, *IEEE transactions on knowledge and data engineering* 24 (1) (2012) 30–44 (2012).
- [81] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 2818–2826 (2016). doi:10.1109/CVPR.2016.308.
URL <https://doi.org/10.1109/CVPR.2016.308>
- [82] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, Ieee, 2009, pp. 248–255 (2009).
- [83] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. MÅzller, How to explain individual classification decisions, *Journal of Machine Learning Research* 11 (Jun) (2010) 1803–1831 (2010).
- [84] P. M. Rasmussen, T. Schmah, K. H. Madsen, T. E. Lund, S. C. Strother, L. K. Hansen, Visualization of nonlinear classification models in neuroimaging (2012).

- [85] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).
- [86] J. M. Zurada, A. Malinowski, I. Cloete, Sensitivity analysis for minimization of input data dimension for feedforward neural network, in: Circuits and Systems, 1994. ISCAS'94., 1994 IEEE International Symposium on, Vol. 6, IEEE, 1994, pp. 447–450 (1994).
- [87] A. Sung, Ranking importance of input parameters of neural networks, Expert Systems with Applications 15 (3-4) (1998) 405–411 (1998).
- [88] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global sensitivity analysis: the primer, John Wiley & Sons, 2008 (2008).
- [89] L. M. Zintgraf, T. S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: Prediction difference analysis, CoRR abs/1702.04595 (2017). arXiv:1702.04595.
URL <http://arxiv.org/abs/1702.04595>
- [90] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, nature 323 (6088) (1986) 533 (1986).
- [91] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, B. Kim, The (un) reliability of saliency methods, arXiv preprint arXiv:1711.00867 (2017).
- [92] S. Bazen, X. Joutard, The taylor decomposition: A unified generalization of the oaxaca method to nonlinear models (2013).
- [93] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (7) (2015) e0130140 (2015).
- [94] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognition 65 (2017) 211–222 (2017).

- [95] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: *European conference on computer vision*, Springer, 2014, pp. 818–833 (2014).
- [96] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, *arXiv preprint arXiv:1412.6806* (2014).
- [97] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929 (2016).
- [98] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, S. Dähne, Learning how to explain neural networks: Patternnet and patternattribution, *arXiv preprint arXiv:1705.05598* (2017).
- [99] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, S. Dähne, Patternnet and patternlrp—improving the interpretability of neural networks, *stat 1050* (2017) 16 (2017).
- [100] P. Berkes, L. Wiskott, On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields, *Neural computation* 18 (8) (2006) 1868–1895 (2006).
- [101] D. Erhan, Y. Bengio, A. Courville, P. Vincent, Visualizing higher-layer features of a deep network, *University of Montreal* 1341 (3) (2009) 1 (2009).
- [102] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, J. Yosinski, Plug & play generative networks: Conditional iterative generation of images in latent space., in: *CVPR*, Vol. 2, 2017, p. 7 (2017).
- [103] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, in: *Advances in Neural Information Processing Systems*, 2016, pp. 3387–3395 (2016).
- [104] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Advances in neural information processing systems*, 2012, pp. 1097–1105 (2012).

- [105] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579 (2015).
- [106] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, arXiv preprint arXiv:1704.05796 (2017).
- [107] B. Zhou, D. Bau, A. Oliva, A. Torralba, Interpreting deep visual representations via network dissection, *IEEE transactions on pattern analysis and machine intelligence* (2018).
- [108] R. A. Fisher, On the probable error of a coefficient of correlation deduced from a small sample, *Metron* 1 (1921) 3–32 (1921).
- [109] D. C. Montgomery, *Design and analysis of experiments*, John Wiley & sons, 2017 (2017).
- [110] S. R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, *IEEE transactions on systems, man, and cybernetics* 21 (3) (1991) 660–674 (1991).
- [111] S. K. Murthy, Automatic construction of decision trees from data: A multi-disciplinary survey, *Data mining and knowledge discovery* 2 (4) (1998) 345–389 (1998).
- [112] H. Lakkaraju, S. H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, in: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, 2016*, pp. 1675–1684 (2016). doi:10.1145/2939672.2939874.
URL <http://doi.acm.org/10.1145/2939672.2939874>
- [113] R. L. Rivest, Learning decision lists, *Machine Learning* 2 (3) (1987) 229–246 (Nov 1987). doi:10.1023/A:1022607331053.
URL <https://doi.org/10.1023/A:1022607331053>
- [114] J. R. Quinlan, Generating production rules from decision trees., in: *ijcai*, Vol. 87, Citeseer, 1987, pp. 304–307 (1987).

- [115] R. A. Adams, Q. J. Huys, J. P. Roiser, Computational psychiatry: towards a mathematically informed understanding of mental illness, *J Neurol Neurosurg Psychiatry* 87 (1) (2016) 53–63 (2016).
- [116] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: *Proceedings of the 21st ACM International Conference on Multimedia, MM '13*, ACM, New York, NY, USA, 2013, pp. 223–232 (2013). doi:10.1145/2502081.2502282.
URL <http://doi.acm.org/10.1145/2502081.2502282>
- [117] T. Chen, D. Borth, T. Darrell, S.-F. Chang, Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks, *arXiv preprint arXiv:1410.8586* (2014).
- [118] P. Blandfort, D. Patton, W. R. Frey, S. Karaman, S. Bhargava, F.-T. Lee, S. Varia, C. Kedzie, M. B. Gaskell, R. Schifanella, et al., Multimodal social media analysis for gang violence prevention, *arXiv preprint arXiv:1807.08465* (2018).
- [119] M.-T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, *arXiv preprint arXiv:1508.04025* (2015).
- [120] S. Sukhbaatar, J. Weston, R. Fergus, et al., End-to-end memory networks, in: *Advances in neural information processing systems*, 2015, pp. 2440–2448 (2015).
- [121] J. Weston, S. Chopra, A. Bordes, Memory networks, *CoRR abs/1410.3916* (2014). arXiv:1410.3916.
URL <http://arxiv.org/abs/1410.3916>
- [122] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: *International conference on machine learning*, 2015, pp. 2048–2057 (2015).
- [123] Q. Zhang, Y. N. Wu, S.-C. Zhu, Interpretable convolutional neural networks, in: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8827–8836 (2018).

- [124] E. R. Kandel, J. H. Schwartz, T. M. Jessell, D. of Biochemistry, M. B. T. Jessell, S. Siegelbaum, A. Hudspeth, Principles of neural science, Vol. 4, McGraw-hill New York, 2000 (2000).
- [125] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, D. Fox, A joint model of language and perception for grounded attribute learning, arXiv preprint arXiv:1206.6423 (2012).
- [126] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017, pp. 1–10 (2017).
- [127] O. Ishaq, S. K. Sadanandan, C. Whlby, Deep fish: Deep learningbased classification of zebrafish deformation for high-throughput screening, SLAS DISCOVERY: Advancing Life Sciences R&D 22 (1) (2017) 102–107, pMID: 27613194 (2017). arXiv:<https://doi.org/10.1177/1087057116667894>, doi:10.1177/1087057116667894. URL <https://doi.org/10.1177/1087057116667894>
- [128] L. Arras, G. Montavon, K.-R. Müller, W. Samek, Explaining recurrent neural network predictions in sentiment analysis, in: Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2017, pp. 159–168 (2017). URL <http://aclweb.org/anthology/W/W17/W17-5221.pdf>
- [129] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, W. Samek, Analyzing classifiers: Fisher vectors and deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2912–20 (2016). doi:10.1109/CVPR.2016.318. URL <http://dx.doi.org/10.1109/CVPR.2016.318>
- [130] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: Security and Privacy (EuroS&P), 2016 IEEE European Symposium on, IEEE, 2016, pp. 372–387 (2016).

- [131] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami, Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ACM, 2017, pp. 506–519 (2017).
- [132] J. Folz, S. Palacio, J. Hees, D. Borth, A. Dengel, Adversarial defense based on structure-to-signal autoencoders, CoRR abs/1803.07994 (2018). [arXiv:1803.07994](https://arxiv.org/abs/1803.07994).
URL <http://arxiv.org/abs/1803.07994>
- [133] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199 (2013).
- [134] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427–436 (2015).
- [135] Y. LeCun, Y. Bengio, et al., Convolutional networks for images, speech, and time series, The handbook of brain theory and neural networks 3361 (10) (1995) 1995 (1995).
- [136] C. O. S. Sorzano, J. Vargas, A. P. Montano, A survey of dimensionality reduction techniques, arXiv preprint arXiv:1403.2877 (2014).
- [137] A. Gisbrecht, B. Hammer, Data visualization by nonlinear dimensionality reduction, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5 (2) (2015) 51–73 (2015).
- [138] K. Pearson, Liii. on lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11) (1901) 559–572 (1901).
- [139] G. McLachlan, Discriminant analysis and statistical pattern recognition, Vol. 544, John Wiley & Sons, 2004 (2004).
- [140] C. Spearman, ” general intelligence,” objectively determined and measured, The American Journal of Psychology 15 (2) (1904) 201–292 (1904).

- [141] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, E. J. Strahan, Evaluating the use of exploratory factor analysis in psychological research., *Psychological methods* 4 (3) (1999) 272 (1999).
- [142] S. R. Briggs, J. M. Cheek, The role of factor analysis in the development and evaluation of personality scales, *Journal of personality* 54 (1) (1986) 106–148 (1986).
- [143] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, *Journal of machine learning research* 9 (Nov) (2008) 2579–2605 (2008).
- [144] H. Bourlard, Y. Kamp, Auto-association by multilayer perceptrons and singular value decomposition, *Biological cybernetics* 59 (4-5) (1988) 291–294 (1988).
- [145] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, Adversarial autoencoders, *arXiv preprint arXiv:1511.05644* (2015).
- [146] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, E. P. Xing, Toward controlled generation of text, *arXiv preprint arXiv:1703.00955* (2017).
- [147] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, *ACM computing surveys (CSUR)* 41 (3) (2009) 15 (2009).
- [148] S. Tschiatschek, R. Iyer, H. Wei, J. Bilmes, Learning mixtures of sub-modular functions for image collection summarization, in: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS'14*, MIT Press, Cambridge, MA, USA, 2014, pp. 1413–1421 (2014).
URL <http://dl.acm.org/citation.cfm?id=2968826.2968984>
- [149] Y. Zhao, R. Hong, J. Jiang, Visual summarization of image collections by fast ransac, *Neurocomputing* 172 (2016) 48 – 52 (2016).
doi:<https://doi.org/10.1016/j.neucom.2014.09.095>.
URL <http://www.sciencedirect.com/science/article/pii/S0925231215005986>
- [150] M. A. Fischler, R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM* 24 (6) (1981) 381–395 (1981).

- [151] B. J. Frey, D. Dueck, Clustering by passing messages between data points, *science* 315 (5814) (2007) 972–976 (2007).
- [152] Z. R. Samani, M. E. Moghaddam, A knowledge-based semantic approach for image collection summarization, *Multimedia Tools and Applications* 76 (9) (2017) 11917–11939 (May 2017). doi:10.1007/s11042-016-3840-1.
URL <https://doi.org/10.1007/s11042-016-3840-1>
- [153] J. E. Camargo, F. A. González, Multimodal latent topic analysis for image collection summarization, *Information Sciences* 328 (2016) 270–287 (2016).
- [154] A. Jaffe, M. Naaman, T. Tassa, M. Davis, Generating summaries and visualization for large collections of geo-referenced photographs, in: *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, ACM, 2006, pp. 89–98 (2006).
- [155] D. Das, A. F. Martins, A survey on automatic text summarization, *Literature Survey for the Language and Statistics II course at CMU 4* (2007) 192–195 (2007).
- [156] A. Nenkova, K. McKeown, A survey of text summarization techniques, in: *Mining text data*, Springer, 2012, pp. 43–76 (2012).
- [157] K. Kucher, A. Kerren, Text visualization techniques: Taxonomy, visual survey, and community insights, in: *Visualization Symposium (PacificVis)*, 2015 IEEE Pacific, IEEE, 2015, pp. 117–121 (2015).
- [158] M. Gambhir, V. Gupta, Recent automatic text summarization techniques: a survey, *Artificial Intelligence Review* 47 (1) (2017) 1–66 (2017).
- [159] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, K. Kochut, Text summarization techniques: a brief survey, *arXiv preprint arXiv:1707.02268* (2017).
- [160] D. Wang, S. Zhu, T. Li, Y. Gong, Comparative document summarization via discriminative sentence selection, *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6 (3) (2012) 12 (2012).

- [161] F. Horn, L. Arras, G. Montavon, K. Müller, W. Samek, Exploring text datasets by visualizing relevant words, CoRR abs/1707.05261 (2017). arXiv:1707.05261.
URL <http://arxiv.org/abs/1707.05261>
- [162] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, Applied statistics (1979) 20–28 (1979).
- [163] L. Rokach, Ensemble-based classifiers, Artificial Intelligence Review 33 (1-2) (2010) 1–39 (2010).
- [164] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, in: Mining text data, Springer, 2012, pp. 415–463 (2012).
- [165] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowledge-Based Systems 89 (2015) 14–46 (2015).
- [166] Y. Fang, L. Si, N. Somasundaram, Z. Yu, Mining contrastive opinions on political texts using cross-perspective topic model, in: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, ACM, New York, NY, USA, 2012, pp. 63–72 (2012). doi:10.1145/2124295.2124306.
URL <http://doi.acm.org/10.1145/2124295.2124306>
- [167] D. Vilares, Y. He, Detecting perspectives in political debates, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1573–1582 (2017).
- [168] G. Li, R. Law, H. Q. Vu, J. Rong, X. R. Zhao, Identifying emerging hotel preferences using emerging pattern mining technique, Tourism management 46 (2015) 311–321 (2015).
- [169] R. Andrews, J. Diederich, A. B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-based systems 8 (6) (1995) 373–389 (1995).
- [170] R. Krishnan, G. Sivakumar, P. Bhattacharya, Extracting decision trees from trained neural networks, Pattern Recognition 32 (12) (1999) 1999 – 2009 (1999). doi:[https://doi.org/10.1016/S0031-3200\(99\)00100-0](https://doi.org/10.1016/S0031-3200(99)00100-0)

[//doi.org/10.1016/S0031-3203\(98\)00181-2](https://doi.org/10.1016/S0031-3203(98)00181-2).

URL <http://www.sciencedirect.com/science/article/pii/S0031320398001812>

- [171] O. Boz, Extracting decision trees from trained neural networks, in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2002, pp. 456–461 (2002).
- [172] H. Chipman, E. George, R. McCulloh, Making sense of a forest of trees, *Computing Science and Statistics* (1998) 84–92 (1998).
- [173] H. F. Tan, G. Hooker, M. T. Wells, Tree space prototypes: Another look at making tree ensembles interpretable, arXiv preprint arXiv:1611.07115 (2016).
- [174] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).
- [175] A. Kovashka, O. Russakovsky, L. Fei-Fei, K. Grauman, et al., Crowdsourcing in computer vision, *Foundations and Trends® in Computer Graphics and Vision* 10 (3) (2016) 177–243 (2016).
- [176] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, *Journal of Machine Learning Research* 11 (Apr) (2010) 1297–1322 (2010).
- [177] F. Rodrigues, F. Pereira, B. Ribeiro, Learning from multiple annotators: distinguishing good from random labelers, *Pattern Recognition Letters* 34 (12) (2013) 1428–1436 (2013).
- [178] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, J. Dy, Modeling annotator expertise: Learning when everybody knows a bit of something, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 932–939 (2010).
- [179] D. Patton, P. Blandfort, W. Frey, M. Gaskell, S. Karaman, Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators, in:

- Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019 (2019).
- [180] C. Eickhoff, Cognitive biases in crowdsourcing, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, ACM, 2018, pp. 162–170 (2018).
- [181] F. V. Waugh, Regressions between sets of variables, *Econometrica, Journal of the Econometric Society* (1942) 290–310 (1942).
- [182] C. E. Hopkins, Statistical analysis by canonical correlation: a computer application., *Health services research* 4 (4) (1969) 304 (1969).
- [183] C. Creighton, S. Hanash, Mining gene expression databases for association rules, *Bioinformatics* 19 (1) (2003) 79–86 (2003).
- [184] C. Ordonez, N. Ezquerra, C. A. Santana, Constraining and summarizing association rules in medical data, *Knowledge and information systems* 9 (3) (2006) 1–2 (2006).
- [185] R. Sherhod, P. N. Judson, T. Hanser, J. D. Vessey, S. J. Webb, V. J. Gillet, Emerging pattern mining to aid toxicological knowledge discovery, *Journal of chemical information and modeling* 54 (7) (2014) 1864–1879 (2014).
- [186] P. R. Montague, R. J. Dolan, K. J. Friston, P. Dayan, Computational psychiatry, *Trends in Cognitive Sciences* 16 (1) (2012) 72 – 80, special Issue: Cognition in Neuropsychiatric Disorders (2012). doi:<https://doi.org/10.1016/j.tics.2011.11.018>. URL <http://www.sciencedirect.com/science/article/pii/S1364661311002518>
- [187] K. E. Stephan, C. Mathys, Computational approaches to psychiatry, *Current opinion in neurobiology* 25 (2014) 85–92 (2014).
- [188] Y. Kawai, Y. Oshima, Y. Sasamoto, Y. Nagai, M. Asada, A computational model for child inferences of word meanings via syntactic categories for different ages and languages, *IEEE Transactions on Cognitive and Developmental Systems* (2018).

- [189] M. Imai, L. Li, E. Haryu, H. Okada, K. Hirsh-Pasek, R. M. Golinkoff, J. Shigematsu, Novel noun and verb learning in chinese-, english-, and japanese-speaking children, *Child development* 79 (4) (2008) 979–1000 (2008).

Fusion Strategies for Learning User Embeddings with Neural Networks

Philipp Blandfort^{1,2}, Tushar Karayil¹, Federico Raue¹, Jörn Hees¹ and Andreas Dengel^{1,2}
 email: `firstname.lastname@dfki.de`

Abstract—Growing amounts of online user data motivate the need for automated processing techniques. In case of user ratings, one interesting option is to use neural networks for learning to predict ratings given an item and a user. While training for prediction, such an approach at the same time learns to map each user to a vector, a so-called user embedding. Such embeddings can for example be valuable for estimating user similarity. However, there are various ways how item and user information can be combined in neural networks, and it is unclear how the way of combining affects the resulting embeddings.

In this paper, we run an experiment on movie ratings data, where we analyze the effect on embedding quality caused by several fusion strategies in neural networks. For evaluating embedding quality, we propose a novel measure, Pair-Distance Correlation, which quantifies the condition that similar users should have similar embedding vectors. We find that the fusion strategy affects results in terms of both prediction performance and embedding quality. Surprisingly, we find that prediction performance not necessarily reflects embedding quality. This suggests that if embeddings are of interest, the common tendency to select models based on their prediction ability should be reconsidered.

I. INTRODUCTION

The past two decades have seen an exponential proliferation of user-generated content across the Internet, including social media posts, user activities and ratings. Such user data has been used in a variety of ways. Examples include the detection of users’ sentiment from product reviews [1], but user data has also been used to train models for predicting where users will click [2] or which items they will like [3]. Such detection and prediction tasks typically have direct practical motivations. It can, however, be important as well to add an explanatory component to such detection and prediction systems. In particular, this importance can be due to legal reasons, since the European legislation (GDPR [4]) now grants users the right to ask for a simple explanation of any automatic decisions that affect them. There are several possibilities for combining analysis with detection or prediction, in order to make AI systems more understandable.

One way, for example, is to build on understandable mid-level concepts for detection, such that the trained model automatically has an explanatory quality. This approach is adopted in aspect-based sentiment detection (e.g., [5]), which is commonly applied to user reviews to not only detect the overall sentiment towards a product but simultaneously describe which aspects of the product are responsible for

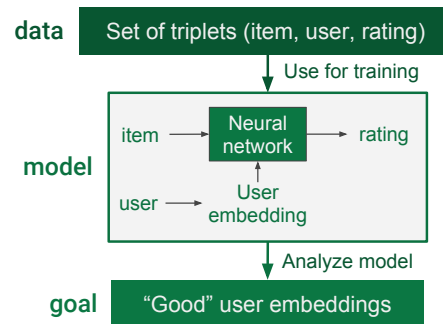


Fig. 1: Illustration of the neural-network-based approach adopted in this paper for learning user embeddings. A neural network learns to fit the data while simultaneously embedding the users, based on item ratings from different users. Our main questions are: How to quantitatively evaluate the learned embeddings? And what are the effects of fusing item and user information in a particular way?

the user’s opinion. Another approach is to fit the given data with a complex model (often a neural network) and then derive explanations from the trained model by means of sophisticated analysis techniques. This direction includes recent efforts related to heatmapping techniques that are used for explaining decisions of machine learning models (e.g., [6], [7]). For example, in the field of medicine, Sturm et al. [8] show how to train a neural network on classifying EEG data and then use a heatmapping technique to generate explanations for the network’s classification decisions. Yet another case of combining prediction with analysis would be representation learning, where some concept of interest (such as the user) is mapped to a vector as part of a larger model that fits the given data. An example for such an approach is proposed by Amir et al. [9]. In their paper, users are embedded to a vector and then fused into a neural network for predicting textual contents (twitter texts). In this regard, neural networks represent an interesting choice as a model. This is because, in the past few years, many techniques have been proposed for analyzing them (such as the ones mentioned above).

We see that most of the mentioned approaches involve training a neural network on user data, which combines user and item information for prediction. Several strategies exist for such an information fusion, but so far the effect of this choice has barely been analyzed.

Hence, in this work we explore the direction of using

¹German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany

²Technische Universität Kaiserslautern, Germany

neural networks for simultaneously fitting user data and learning vector representations for users, and analyze the effects of different fusion variants. As user data, we decide to use movie ratings from different users, where we represent movies as (dense) feature vectors based on tags. Our goal is to fit this data with a neural network that takes user ID and movie features as input, while predicting the corresponding ratings and learning to embed the users into vectors. The focus thereby lies on the representation learning component, i.e., we want to find out how to learn “good” embeddings. To this end, we mainly address these questions: Which way of combining user and movie information is suitable for such a task? What is the effect of embedding size? And how can we define and quantify the quality of embeddings? In this regard, the contributions of this work can be summed up as following:

- 1) We propose a novel evaluation measure for quality of user embeddings.
- 2) We analyze the effect of various fusion strategies and embedding size on the resulting quality of learned embeddings.

The rest of the paper is organized as follows. Section II summarizes the previous works that have been related to this field of research. Section III formalizes the task in a detailed manner and introduces the proposed measure for embedding quality. Section IV explains the relevant fusion strategies for neural networks. Section V describes the experiment performed. Section VI summarizes our findings and mentions some future directions.

II. RELATED WORK

A. Learning user embeddings

The main goal of this paper is to learn (meaningful) user embeddings. In the literature, user information has been used in various ways. For example, user information can be exploited for adding cognitive information to the model. This was done by Yamagashi et al. [10] who defined a user embedding as context for learning different speaking styles, such as reading, joyful, and sad. There, the user information is defined in terms of two components: phonetic and linguistic. Additionally, user information has been used for detecting sarcasm and mental health conditions based on social media data [11], [9]. In both of these scenarios, neural networks based on `Paragraph2Vec` [12] are trained on textual contents with the goal of learning user-dependent word-usage patterns. In this process, the model automatically learns user embeddings that are based on the relationship between users and their texts.

Both of these papers analyze the learned user embeddings in order to obtain insights about user behavior. The authors, however, do not propose a formal measure for evaluating embedding quality, and the fusion strategy in both cases is simple concatenation. We will propose a novel measure for quantitative evaluation of embedding quality in Section III-D. Also, whether concatenation is the most appropriate fusion strategy for learning embeddings is far from obvious.

Indeed, we can find several other fusion strategies for neural networks in different areas, which we shall briefly discuss now.

B. Fusion strategies in neural networks

Fusion strategies have been applied to two or more modalities for joining representations and predictions. One possibility is to have a shared representation in the model [13]. Furthermore, each modality is learned individually as a first layer and then both components are joined into a shared representation as a second layer. This can be seen as a early fusion. Additionally, the fusion can also be presented in the middle or at the end of the model. For example, a common approach in Visual-Question Answering (VQA) is to first obtain visual and text embeddings after applying a Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), respectively [14]. Then, a simple Hadamard product (i.e. element-wise multiplication) is used as a fusion method in the model. The previous two approaches are based on concatenation or multiplication operations. Another approach is to apply tensor operations to the multimodal embeddings. For example, the MUTAN model [15] factorizes a multimodal tensor generated by the question and image embeddings. Similarly, tensor products are used for information fusion in some works of distributional semantics (e.g., [16], [17], [18]). Some of these works ([16], [18]) even include a systematic comparison of fusion strategies in terms of effects on embeddings of adjective-noun combinations. However, for the case of user embeddings, no such comparative study exists.

III. PRELIMINARIES

A. Problem statement

We assume that we are given data of the form $\{(x_1, u_1, R_1), \dots, (x_m, u_m, R_m)\}$, where $x_i \in \mathbb{R}^n$ are input items (such as movies), $u_i \in \mathcal{U}$ correspond to users with \mathcal{U} the space of user identifiers, and $R_i \in \mathbb{R}$ is the rating which the user u_i assigned to item x_i . The goal is to find a mapping $e : \mathcal{U} \rightarrow \mathbb{R}^z$ that assigns each user to a real-valued vector, which we refer to as the user’s *embedding*. For example, a user with ID `user_1` could be mapped to a 3-D vector $e(\text{user}_1) = [0.2, 0.1, 0.7]^T$, where $z = 3$. We require that these embeddings are “meaningful” in the sense that similarity of embeddings should reflect similarity between users. Intuitively, we want to represent the users such that it is easy for us to see how similar they are in terms of how they rate items. (This part will be formalized as a novel measure in Section III-D.)

In this paper, we analyze how this goal can be achieved by fitting the given data with a neural network that simultaneously learns embeddings for the users. The main questions we address are: How does embedding size z relate to the quality of the learned embeddings and the ability to fit the data? And, since such a neural network needs to combine input item and user information for predicting a rating, which fusion strategy is most appropriate for this task?

B. Functional Data Analysis (FDA)

In order to foster a deeper understanding of the problem, we describe its relation to a particular branch of mathematics, namely Functional Data Analysis (FDA). This will show how learning user embeddings can be understood as finding vector representations for functions. This insight will then be useful later for seeing how embeddings can be evaluated quantitatively.

Mathematically, we can model the data generation process analogously to the modeling in FDA (compare, e.g., with the description in the survey of Jacques and Preda [19]): We assume that there is a functional random variable

$$F : \Omega \rightarrow \{f : I \rightarrow \mathbb{R}\},$$

i.e., F is a random variable which has functions (from I to \mathbb{R}) as values. Any such function f describes a particular way of rating items, and corresponds to a single user. Now, a set of observations $\{f_1, \dots, f_l\}$ of F is referred to as *functional data*. In practice, these rating functions are not given directly, but instead, for each function f_i a set of samples $\{(x_{i;1}, f_i(x_{i;1})), \dots, (x_{i;m_i}, f_i(x_{i;m_i}))\}$ is provided. We can put all this information together into a set with elements of the form $(i, x_{i;j}, f_i(x_{i;j}))$, which shows the equivalence to the rating data introduced in our problem statement (Section III-A). So our goal of learning user embeddings essentially means that we are trying to find vector representations of functions based on lists of samples. This corresponds to dimensionality reduction of functions, which is a sub-task of FDA [20].

This view of the problem should make another thing clear: User embeddings are representations of the users' rating functions. There is no a priori justification for assuming that any other properties of the users (apart from their rating behaviors) would be incorporated into user embeddings by fitting such data. Hence, relations between user attributes (such as gender or location) and embeddings can be useful to analyze the role of such attributes for user behavior, but are only suitable for evaluating the embeddings if there is a known connection between rating behavior and the given user attributes.

C. User similarity

In Section III-A we formulated the goal that similarities of embeddings should reflect similarities of the corresponding users (or, to be more precise, their rating behaviors, as we have just argued in the previous section). Before we can turn this criterion into a quantitative measure, we need to quantify similarity of users.

In the collaborative filtering literature, we find several proposed methods for computing such similarities, including Pearson correlation, Spearman correlation, cosine vector similarity, adjusted cosine vector similarity, and mean-squared difference [21]. Out of these common options we choose to estimate similarity based on the mean-squared difference of their ratings. Our choice has two main reasons: First, mean-squared difference is most similar to the L2 distance between functions, which is a common measure used in FDA

for computing distance between functions [19]. Second, the mean-square difference of user ratings is interpretable as expected value (assuming uniform prior over input items) of the squared difference of their ratings for the same item.

D. Pair-Distance Correlation measure (PDC)

We are not aware of any existing measure for evaluating embeddings based on function similarity. Thus, for measuring the quality of learned embeddings, we introduce a novel performance measure, which we call *Pair-Distance Correlation* (PDC) measure. Two crucial elements of the presented measure are the distance metric d_E on the embedding space, and the distance measure d_U on the user space.

The choice for these two elements should be informed by the purpose of learning user embeddings. For interpretability, it makes sense to consider intuitive distances on the embedding space (d_E) such as the Euclidean distance. As mentioned above, we generally consider the mean-squared difference to be an appropriate choice for d_U , but in certain scenarios one might want to deviate from that (e.g., mean-absolute difference if more weight should be given to small differences). Once d_E and d_U are chosen, the PDC of an embedding function is computed as follows:

Algorithm 1 Algorithm for computing Pair-Distance Correlation

Input: set $\{(x_1, u_1, R_1), \dots, (x_n, u_n, R_n)\}$ with input items $x_i \in \mathbb{R}^n$, user identifiers $u_i \in \mathcal{U}$ and ratings $R_i \in \mathbb{R}$; embedding function $e : \mathcal{U} \rightarrow \mathbb{R}^z$; distance measures d_E on \mathbb{R}^z and d_U on \mathcal{U} ; threshold $t \in \mathbb{N}_{>0}$

Output: PDC score of e with respect to d_E and d_U

- 1: set $l_E = \text{list}()$, $l_U = \text{list}()$
 - 2: **for** all users $u_i, u_j \in \mathcal{U}$ with at least t items rated by both **do**
 - 3: based on all items rated by both u_i and u_j , compute $d_U(u_i, u_j)$ and append it to l_U
 - 4: compute $d_E(e(u_i), e(u_j))$ and append it to l_E
 - 5: **end for**
 - 6: **return** Pearson correlation coefficient between l_E and l_U
-

Being based on Pearson correlation, the resulting score takes values in $[-1, 1]$, where higher values are preferable and 1 is the best possible outcome. Note that random embeddings can be expected to achieve a PDC score around 0. A high PDC measure (close to 1) means, that in general if a user embedding $e(u_i)$ is more similar to $e(u_j)$ than $e(u_k)$ with respect to d_E (i.e., $d_E(e(u_i), e(u_j)) < d_E(e(u_i), e(u_k))$), then the rating behavior of user u_i is more similar to the behavior of u_j than to that of u_k (in terms of the similarity measure explained in Section III-C, i.e., $d_U(u_i, u_j) < d_U(u_i, u_k)$). In other words, PDC evaluates whether an embedding function preserves distance relations.

IV. FUSION STRATEGIES

We would like to train a neural network that takes item information x as input and user embedding e as additional

context signal for predicting the user’s rating. In general we distinguish between three ways for incorporating such context information into neural networks:

- 1) *neuron-level fusion*: based on context signal alter hidden states at some layer
- 2) *weight-level fusion*: based on context signal alter weights of some layer
- 3) combinations of the former two

We outline three specific approaches that fall into the first two of these categories, which we will also use in our experiments later on. Then, we will also have a closer look at Factorization Machines (FMs) [22], which have shown top performance for various tasks involving user-dependent prediction [23], [24], [25]. The paper that introduces FMs [22] also explains how FMs can mimic many other popular recommendation system methods, including matrix factorization and specialized methods such as SVD++ [26] or PITF [27]. Note that even though FM is not a neural network method per se, it can, after a small modification, be understood as weight-level fusion approach.

A. Neuron-level fusion

We focus on mask-based methods for neuron-level fusion. In mask-based methods, a mask of the same size as the hidden state at some level is computed from the context signal and then combined with the hidden state in an element-wise manner for an update. Two common fusion approaches are considered:

- For using *additive masks* (Add) on any hidden state x , we compute a mask of the shape of x by multiplying the context vector with a weight matrix of suitable shape, and then add this mask to the original state.
- *Multiplicative masks* (Mul) work analogously but combine mask and hidden state by element-wise multiplication (i.e., Hadamard product).

Note that additive masks are equivalent to concatenation of input $x \in \mathbb{R}^n$ and context signal $e \in \mathbb{R}^z$ one layer earlier, since it holds that

$$y = W \begin{bmatrix} x \\ e \end{bmatrix} = [W_1 | W_2] \begin{bmatrix} x \\ e \end{bmatrix} = W_1 x + W_2 e, \quad (1)$$

where $\begin{bmatrix} x \\ e \end{bmatrix}$ stands for the concatenation of x and e and the $m \times (n + z)$ -matrix W is split into the $m \times n$ -matrix W_1 and the $m \times z$ matrix W_2 .

B. Weight-level fusion

We describe *tensor fusion* as one way to make weights context-dependent. We made this choice because tensor fusion is a basic approach which we find very suitable for illustrating the general principles of weight-level fusion. Other approaches such as the one inspired by Singular Value Decomposition (introduced for one-shot learning in [28]) can typically be understood as a modification of tensor fusion. Different from neuron-level fusion, which can normally be applied in exactly the same way in linear layers and convolutional layers, the details of weight-level fusion depend on the type of layer.

We describe tensor fusion on a linear layer. So let us assume that we have input $x \in \mathbb{R}^n$, context $e \in \mathbb{R}^z$ and want to map this to the output space \mathbb{R}^m . The standard output of such a linear layer that ignores all context information is then

$$\text{fc}(x) := b + Wx, \quad (2)$$

where $W \in \mathbb{R}^{m \times n}$ is a weight matrix and $b \in \mathbb{R}^m$ a bias term. The basic idea of tensor fusion is to make the weight matrix W dependent on the context e by adding a context-dependent part to it. More precisely, we define $W(e) := W + eT$ where $T \in \mathbb{R}^{z \times m \times n}$ is a third-order tensor, and eT is calculated as $eT = \sum_{i=1}^z e_i T_{i, \cdot, \cdot}$. The final output of the linear tensor fusion layer is then given by

$$\text{tensor}(x, e) := b + W(e)x = b + (W + eT)x \quad (3)$$

$$= b + Wx + eTx \quad (4)$$

C. Factorization Machines

Using a slightly different notation than in the original paper [22], we can write the model equation of a FM (of degree 2 and rank z) as follows:

$$\text{FM}(x) := b + Wx + \sum_{i=1}^z \sum_{j=i+1}^n x_i V_i \cdot V_j^T \cdot x_j, \quad (5)$$

where $x \in \mathbb{R}^n$, $W \in \mathbb{R}^n$, $V \in \mathbb{R}^{n \times z}$. (Note that the output dimension m is 1 for FMs.)

A modified version of FM turns out to be a special case of tensor fusion, which we will now explain. By changing the sum over j to go from 1 to n (instead of $i + 1$ to n), we get:

$$\text{FM}_T(x) := b + Wx + \sum_{i=1}^z \sum_{j=1}^n x_i V_i \cdot V_j^T \cdot x_j \quad (6)$$

$$= b + Wx + xVV^T x \quad (7)$$

There are two ways how FM_T can be understood as tensor fusion: First, if we define $T := VV^T \in \mathbb{R}^{n \times n}$, we see that

$$\text{FM}_T(x) = b + Wx + xTx \quad (8)$$

becomes equivalent to tensor fusion that uses the same vector as input and context, and factorizes the weight tensor. Second, we can consider V as embedding matrix, so that x is used as input and its embedding xV as context:

$$\text{FM}_T(x) = b + Wx + (xV)V^T x \quad (9)$$

In this case, V^T takes the role of the tensor T , and we have tensor fusion that shares weights with the embeddings.

We would like to point out that by interpreting the modified FM (which can still capture higher-order dynamics) in any of these two ways, it becomes simple to see how structures similar to FM can be incorporated anywhere into a (potentially large) neural network. In particular, these interpretations as tensor fusion explain how the method can be adapted for higher-dimensional output (where both of the two interpretations we discussed lead to slightly different adaptations).

V. EXPERIMENT

The experiment aims to evaluate the effects of embedding size and fusion strategy on the quality of user embeddings. Apart from the baselines, we also conduct similar experiments with Factorization Machines as a benchmark. The experiment uses the MovieLens-100k dataset [29], which contains 100,000 movie ratings. The neural networks used in this experiment are based on linear layers.

A. Task

We use the MovieLens-100k dataset, which consists of 100,000 movie ratings (1-5) from 943 users and of 1682 movies. Each movie in this dataset has a unique ID and meta-information about title, year of appearance and genre(s). None of these seem overly interesting to use as interpretation input, hence we took the movie genome information from the MovieLens-20M dataset [29] in order to obtain a 1128-dimensional tag-based feature representation of the movies.¹

We train various neural networks on the task of movie rating prediction, given the movie as tag feature vector and the user ID as input. (See task illustration in Figure 2.) It

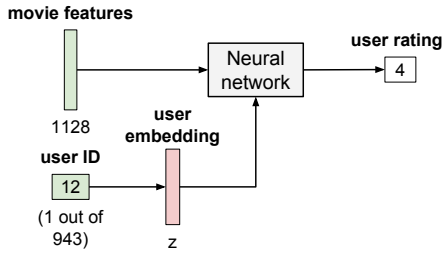


Fig. 2: Illustration of the MovieLens-100k task. A neural network is trained to take movie features as input and user ID as context signal for predicting user ratings. The particular neural network architecture and embedding size are varied in the experiment. Note that everything is trained end-to-end, which includes learning the embeddings.

is important to recall that our main interest lies in learning meaningful embeddings. Embedding quality is evaluated by computing the PDC measure with respect to mean-square difference on users (estimated based on the test data) and Euclidean distance on the embeddings (as introduced in Section III-D). Additionally, we evaluate prediction performance, using the standard recommendation system measures mean average error (MAE) and root mean squared error (RMSE).

B. Architectures

The neural network architectures used for this experiment are illustrated in Figure 3. All of these architectures are based on one or two linear layers, and incorporate the user information by additive masks, multiplicative masks or tensor

fusion, respectively (see Section IV). We deploy all masking mechanisms before applying the activation function. For each of the four fusion methods, we vary the embedding size (2, 4, 8, 16, 32 and 64).

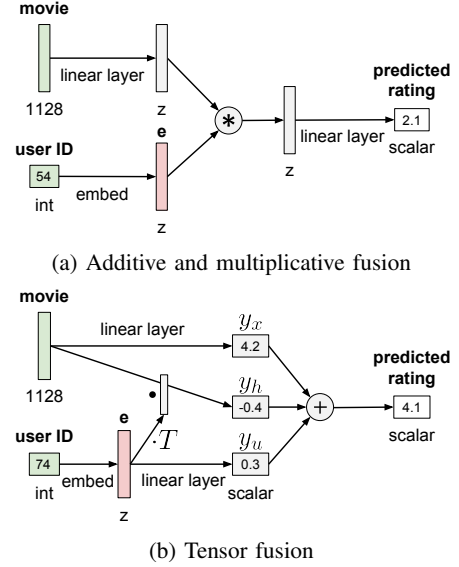


Fig. 3: Illustration of neural network architectures used for the MovieLens-100k experiment. For additive and multiplicative fusion, the basic architecture can be understood as multi-layer perceptron with one hidden layer of size z , where the hidden activations are modified depending on the user signal by means of element-wise addition/multiplication (taking the place of “*”). As tensor fusion approach, we choose a single linear tensor fusion layer, that uses the user embeddings as context signal.

We compare these neural networks against a Factorization Machine (FM). For each user, an embedding in case of FM is obtained by appending the user bias (as learned by W) to the row of the weight matrix V which corresponds to the user.

C. Results

Results can be found in Table I. As two baselines, we include a model that outputs the average rating of the given user and ignores the movie features (user-bias), and a model that adjusts this score based on average user ratings by adding a learned linear combination of the movie features (linear). For the baselines, the (scalar) user-dependent biases were used as embeddings. All reported results are averages of 5-fold cross validation, using the official dataset split. In Figure 4, we additionally show how embedding scores vary across fusion strategies and embedding sizes, depending on the threshold of common items that is used for computing the PDC score. We optimized hyper-parameters (learning rate, number of epochs, and in case of FM also the regularization terms) based on a different split. We implemented all baselines and neural network models in TensorFlow [30], and used Sacred [31] for managing our experiments. FMs

¹Note that for this we had to link the movie IDs between these two datasets, which we did based on movie titles and years of appearance. We dropped the (<200) movies for which no corresponding movie was found.

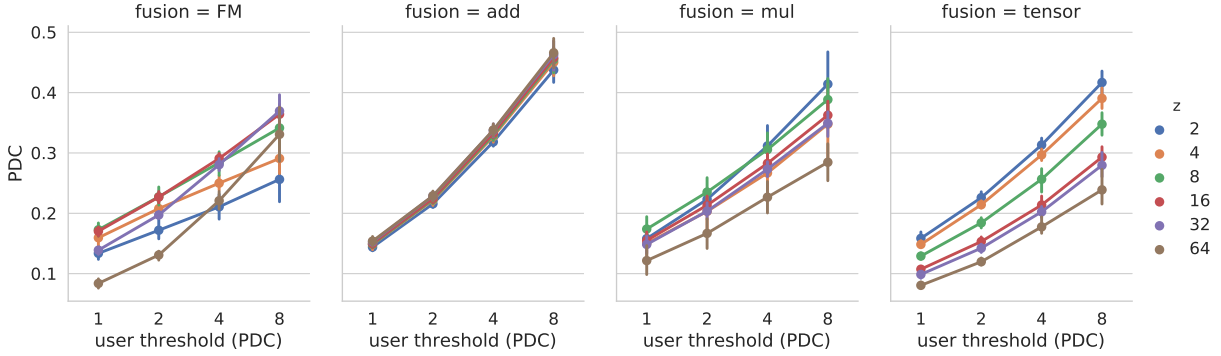


Fig. 4: PDC scores for embeddings of different fusion strategies for various embedding sizes. The threshold that is varied along the x-axis is a parameter of the PDC measure: For computing the PDC measure, only user pairs with this number of common movies are considered. Mean values and standard deviation are calculated based on training and evaluation of 5 folds. Corresponding scores of the linear baseline are 0.12, 0.19, 0.29, 0.42 (thresholds 1, 2, 4, and 8, respectively).

were trained and evaluated in a separate script, using Bayer’s fastFM implementation [25].

D. Analysis

In Table I we see that fusion with multiplicative masks of a moderate embedding size (around 16) works best for prediction. Tensor fusion with similar embedding sizes (16, 32) achieves comparable prediction results. Additive masks yield comparatively poor prediction performances, but achieve higher embedding scores. Both of these parts are largely independent of the chosen embedding size in case of addition, while we can observe a slight trend of increasing embedding quality with growing embedding size. For tensor fusion on the other hand, embedding quality is overall lower, and heavily depends on the chosen embedding size, while lower dimensions yield higher quality. Interestingly, tensor fusion favors high embedding sizes for prediction, which means that these two types of performances are found to be anti-correlated. Multiplicative fusion shows similar trends to those of tensor fusion, so it also prefers low embedding size for learning embeddings, but this effect is somewhat less pronounced. Factorization Machines are highly competitive for prediction when using small embedding sizes, but at around $z = 16$ start overfitting quite heavily. Embedding scores are comparable to those of the multiplicative fusion model. Again, optimal embedding size is different for prediction as compared to embedding quality (8 vs 2 or 4). In general it is surprising that there seems to be no clear relation between prediction performance and embedding quality (compare, e.g., addition with $z = 32$, multiplication with $z = 4$ and tensor fusion with $z = 64$).

Results for PDC score in Table I are all based on a user threshold of 4, i.e., only pairs of users with at least 4 commonly-rated movies were considered for the calculation. In Figure 4 we can see how embedding scores change if we vary this user threshold. Note that for lower thresholds we have many more user pairs to consider but also much more noise in the data, which explains why scores are generally

much lower for lower user thresholds. Most of the effects we discussed above are insensitive to this user threshold we chose for computing the PDC score. One exception to this insensitivity is the observation that for low user thresholds, addition is on par with some tensor fusion models, and even slightly outperformed by certain variants of FM and multiplicative fusion ($z = 8$). Together with the corresponding prediction performances, this suggests that additive fusion generally focuses on less complex user-dependent effects (as compared to FM and multiplicative fusion). Figure 4 also reveals a high standard deviation of the embedding quality for multiplicative fusion. This is generally an undesired property but could indicate that multiplicative fusion might benefit from additional regularization techniques.

Overall, in this experiment additive fusion appears to be a robust choice for learning high quality embeddings, as long as prediction performance is not important.

E. Input sensitivities and user clustering

The model-based approach we adopted for fitting user data while simultaneously learning to represent the users as embedding vectors gives us one other interesting option, which we have not yet mentioned: For any user embedding, the trained model describes the associated rating behavior, which we can analyze further. In particular, for a given user, we can compute input sensitivities (partial derivatives of rating score with respect to input features) in order to find out, which features of a movie make the user more likely to rate the movie higher, and which features the user does not like. Of course, we should not assume that any of our models perfectly fits the true rating behavior of any user, especially in terms of more complex properties such as input sensitivities. Still, if the models achieve reasonable prediction performance, there is good reason to believe that at least some properties are captured correctly. And choosing a model with a simple structure furthermore reduces the chance of ending up with complex statistical artifacts.

TABLE I: Prediction (MAE, RMSE) and embedding scores (PDC) of the presented fusion strategies and baselines. PDC scores are computed with respect to mean-square difference between users and Euclidean distance on embeddings. For calculating the PDC scores, only user pairs with at least 4 common ratings were considered (threshold 4). The +1 in the z column represents the bias term.

Approach	Fusion z	Prediction		PDC (threshold 4)	Params
		MAE	RMSE		
user-bias	0+1	0.87	1.06	0.26	946
linear	0+1	0.76	0.95	0.29	3015
FM	2+1	0.71	0.91	0.26	6214
	4+1	0.71	0.91	0.29	10356
	8+1	0.72	0.92	0.31	18640
	16+1	0.75	0.96	0.29	35208
	32+1	0.80	1.03	0.28	68344
	64+1	0.83	1.07	0.24	134616
add	2	0.74	0.94	0.32	4147
	4	0.74	0.93	0.33	8293
	8	0.73	0.93	0.33	16585
	16	0.73	0.93	0.33	33169
	32	0.74	0.93	0.34	66337
	64	0.74	0.94	0.34	132673
mul	2	0.73	0.92	0.31	4147
	4	0.71	0.91	0.27	8293
	8	0.70	0.90	0.31	16585
	16	0.70	0.90	0.28	33169
	32	0.70	0.90	0.27	66337
	64	0.71	0.90	0.23	132673
tensor	2	0.73	0.92	0.31	5273
	4	0.72	0.91	0.30	9417
	8	0.71	0.91	0.26	17705
	16	0.71	0.90	0.21	34281
	32	0.71	0.90	0.20	67433
	64	0.71	0.91	0.18	133737

Among our neural network architectures, the structure of the tensor fusion model is particularly simple, which even allows for direct interpretation of the learned weights (see Figure 3). The model has a user-independent rating prediction y_x based on the movie features alone. To this baseline prediction, two numbers are added that both depend on the user. The first number is a general user bias y_u , computed from the user embedding. For the second number y_h , the user embedding is mapped to a vector which then serves as weights to compute another linear combination of the movie features. This second weight vector directly describes user-dependent changes in input sensitivity, since there is no non-linearity in the model. Hence, for any user embedding we obtain a corresponding bias term and changes in input sensitivities without having to put any item data through the model.

This looks very different for multiplicative or additive fusion, where user-dependent effects on input sensitivities can vary across items. Input sensitivities (or relevancies) can still be analyzed in this case by using heatmapping techniques (e.g., [6], [7]), but here the risk of observing statistical artifacts becomes higher (since the possibility that input sensitivities can differ across items drastically increases the number of effects to analyze).

To at least get an intuition of how helpful such input

sensitivities might be in practice, we run another experiment as preliminary analysis. In this experiment, we select the tensor fusion model with embedding size 4 (of best performing training fold) and have a closer look at what the model has learned. To this end, we run k-means clustering with 20 clusters on the learned user embeddings. For 3 random clusters, we pick the centroid embedding and read the associated biases and changes in input sensitivities from the model. We also include the highest- and lowest-ranked movies based on these values. The results can be found in Table II. The constellation of features and movies in these results seems coherent and suggests that this is a promising direction for further investigation.

VI. CONCLUSION

In this paper, we introduced the PDC measure for evaluating user embeddings based on similarities of their rating behavior. This novel measure formalizes the intuitive requirement that similar users should be mapped to similar vectors.

We conducted an experiment on movie rating data, where we compared additive, multiplicative, and tensor fusion in neural networks that learn to fit this data while forming vector representations of all the users. In our experiment we found that the fusion strategy has a significant effect on prediction as well as the quality of the learned embeddings. The effect of embedding size on prediction performance and embedding quality seems to largely depend on the chosen fusion strategy. Additive conditioning was mostly unaffected by changes in embedding size and other methods generally favored small embedding sizes for high embedding quality. Surprisingly, good prediction performance does not necessarily reflect the quality of the learned embeddings. In case of tensor fusion, we even observed these two aspects to be anti-correlated.

This is an important finding since one tends to select models based on their prediction ability, but apparently it is not at all clear how well this measure correlates with other aspects of interest, such as “meaningfulness” of embeddings or learned input sensitivities. In our opinion, this finding suggests that much more work is necessary to better understand the internal dynamics of neural networks, especially when fusion of different information is involved and the models are to be used for data analysis.

Finally, it shall be mentioned that, although we formulated the problem in terms of user ratings, the same modeling can directly be applied to other data such as dialogues. In fact, our chosen approach for learning user embeddings fits the theoretical framework of interpretation analysis proposed by [32], and can be seen as a case of model-based interpretation analysis.

ACKNOWLEDGMENT

This work was supported by the BMBF project DeFuseNN (Grant 01IW17002) and the NVIDIA AI Lab (NVAIL) program. Furthermore, the first author received financial support from the Center for Cognitive Science, Kaiserslautern, Germany.

TABLE II: Biases, favorite and least liked movie features and movies associated with centroids of 3 random user clusters. Clustering was done on user embeddings learned by the tensor fusion model with embedding size 4. The biases and scores of movie features were read from the same model, which was also used for ranking the movies. The table contains the 5 highest/lowest ranked features and 3 highest/lowest ranked movies, respectively.

Cluster No.	General bias	Highest ranked		Lowest ranked	
		Movie features	Movies	Movie features	Movies
1	-0.019	italy, unlikeable characters, road trip, character study, parody	Pulp Fiction (1994), A Clockwork Orange (1971), The Big Lebowski (1998)	women, nudity, history, marx brothers, great	Between the Folds (2008), Duma (2005), McFarland USA (2015)
2	-0.033	visuals, dark humor, classic, non-linear, sarcasm	Pulp Fiction (1994), Reservoir Dogs (1992), Taxi Driver (1976)	predictable, childhood, betrayal, cheating, bad acting	You've Got Mail (1998), Ghost (1990), Runaway Bride (1999)
3	0.017	intimate, visually appealing, costume drama, women, whimsical	Cries and Whispers (1972), Last Life in the Universe (2003), Submarino (2010)	chase, 70mm, snakes, tele- portation, chris tucker	Independence Day (1996), Transformers (2007), Men in Black (1997)

REFERENCES

- [1] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10760–10773, 2009.
- [2] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich, "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine," Omnipress, 2010.
- [3] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: introduction and challenges," in *Recommender systems handbook*, pp. 1–34, Springer, 2015.
- [4] EU Council, "EU Regulation 2016/679 General Data Protection Regulation (GDPR)," *Official Journal of the European Union*, vol. 59, pp. 1–88, 2016.
- [5] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, pp. 415–463, Springer, 2012.
- [6] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1 – 15, 2018.
- [7] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, vol. 1, no. 1, pp. 39–48, 2018.
- [8] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of neuroscience methods*, vol. 274, pp. 141–145, 2016.
- [9] S. Amir, G. Coppersmith, P. Carvalho, M. J. Silva, and B. C. Wallace, "Quantifying mental health from social media with neural user embeddings," *arXiv preprint arXiv:1705.00335*, 2017.
- [10] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for hmm-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, pp. 1–5, IEEE, 2004.
- [11] S. Amir, B. C. Wallace, H. Lyu, and P. C. M. J. Silva, "Modelling context with user embeddings for sarcasm detection in social media," *arXiv preprint arXiv:1607.00976*, 2016.
- [12] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, pp. 1188–1196, 2014.
- [13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.
- [14] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [15] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comp. Vis.*, vol. 3, 2017.
- [16] E. Guevara, "A regression model of adjective-noun compositionality in distributional semantics," in *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pp. 33–37, Association for Computational Linguistics, 2010.
- [17] D. Bamman, C. Dyer, and N. A. Smith, "Distributed representations of geographically situated language," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 828–834, 2014.
- [18] M. Hartung, F. Kaupmann, S. Jebbara, and P. Cimiano, "Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, pp. 54–64, 2017.
- [19] J. Jacques and C. Preda, "Functional data clustering: a survey," *Advances in Data Analysis and Classification*, vol. 8, no. 3, pp. 231–255, 2014.
- [20] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, "Functional data analysis," *Annual Review of Statistics and Its Application*, vol. 3, pp. 257–295, 2016.
- [21] S. Gong, "A collaborative filtering recommendation algorithm based on user clustering and item clustering," *JSW*, vol. 5, no. 7, pp. 745–752, 2010.
- [22] S. Rendle, "Factorization machines," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 995–1000, IEEE, 2010.
- [23] S. Rendle, "Social network and click-through prediction with factorization machines," in *KDD-Cup Workshop*, p. 113, 2012.
- [24] I. Bayer and S. Rendle, "Factor models for recommending given names," *ECML PKDD Discovery Challenge*, p. 81, 2013.
- [25] I. Bayer, "fastfm: a library for factorization machines," *arXiv preprint arXiv:1505.00641*, 2015.
- [26] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 426–434, ACM, 2008.
- [27] S. Rendle and L. Schmidt-Thieme, "Pairwise interaction tensor factorization for personalized tag recommendation," in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 81–90, ACM, 2010.
- [28] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Advances in Neural Information Processing Systems*, pp. 523–531, 2016.
- [29] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *Acm transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, p. 19, 2016.
- [30] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: a system for large-scale machine learning," in *OSDI*, vol. 16, pp. 265–283, 2016.
- [31] K. Greff, A. Klein, M. Chovanec, F. Hutter, and J. Schmidhuber, "The sacred infrastructure for computational research," in *Proceedings of the Python in Science Conferences-SciPy Conferences*, 2017.
- [32] P. Blandfort, J. Hees, and D. U. Patton, "An overview of computational approaches for analyzing interpretation," *arXiv preprint arXiv:1811.04028*, 2018.

The Focus-Aspect-Value Model for Explainable Prediction of Subjective Visual Interpretation

Tushar Karayil*
DFKI & TUK
Kaiserslautern, Germany
tushar.karayil@dfki.de

Jörn Hees
DFKI
Kaiserslautern, Germany
joern.hees@dfki.de

Philipp Blandfort
DFKI & TUK
Kaiserslautern, Germany
philipp.blandfort@dfki.de

Andreas Dengel
DFKI & TUK
Kaiserslautern, Germany
andreas.dengel@dfki.de

ABSTRACT

Subjective visual interpretation is a challenging yet important topic in computer vision. Many approaches reduce this problem to the prediction of adjective- or attribute-labels from images. However, most of these do not take attribute semantics into account, or only process the image in a holistic manner. Furthermore, there is a lack of relevant datasets with fine-grained subjective labels. In this paper, we propose the Focus-Aspect-Value (FAV) model to structure the process of capturing subjectivity in image processing, and introduce a novel dataset following this way of modeling. We run experiments on this dataset to compare several deep learning methods and find that incorporating context information based on tensor multiplication outperforms the default way of information fusion (concatenation).

ACM Reference Format:

Tushar Karayil, Philipp Blandfort, Jörn Hees, and Andreas Dengel. 2019. The Focus-Aspect-Value Model for Explainable Prediction of Subjective Visual Interpretation. In *2019 International Conference on Multimedia Retrieval (ICMR '19)*, June 10–13, 2019, Ottawa, ON, Canada. ACM, New York, NY, USA, Article 4, 9 pages. <https://doi.org/10.1145/3323873.3325026>

1 INTRODUCTION

Subjectivity is the phenomenon wherein human perception is influenced by personal feelings, tastes, opinions etc. The variance which arises as a result of this phenomenon plays a crucial role in the visual domain. For example, the meaning that we infer from an image can depend on: our internal templates about the stimuli [30], expectations and learned biases about the visual object [9], context / prior visual input [8], random neural fluctuations in cortex [18] and other factors like personality of the interpreting individual.

*Equal contribution of Karayil and Blandfort.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR '19, June 10–13, 2019, Ottawa, ON, Canada

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6765-3/19/06...\$15.00

<https://doi.org/10.1145/3323873.3325026>

This innate diversity in interpretation has made evaluation and computational modeling of subjectivity a difficult task.

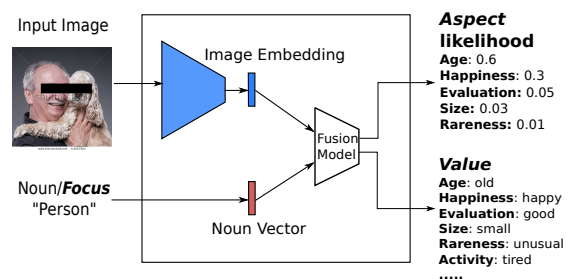


Figure 1: Illustration of the task. The model takes an image and a noun (*focus*) present in the image as input. It outputs the corresponding *aspects* and *value* (of each *aspect*). For the given image in the illustration, since the *focus* is on the noun “person”, the model identifies the *aspects* “age” and “happiness” as the most appropriate. The *value* provided for each *aspect* determines which set of attributes suits the noun in the context of the image. For the *aspect* age in the given example, the *value* output indicates that suitable attributes for the person in the image would be “old”, “elderly”, “mature” or “senior”, in contrast to “young”. The input image is taken from Google’s Conceptual Caption Dataset [29].

The challenge in modeling subjectivity arises from two main sources. First, subjective interpretation by definition is arbitrary in a certain sense, since there is no a priori objective taste, feeling, or opinion, and at times such context information might not be accessible at all. In particular, this poses challenges to evaluation, and in many cases it is reasonable to expect that there will be a larger margin to a perfect score. Second, subjectivity tends to be more fine-grained than objectivity. For example, in images, objectivity that is detected typically is about which entities are visible, while subjective information is rather about characterizing how these entities or the picture as a whole differ from some expectation [8, 9].

To attenuate these issues, previous methods typically consider holistic aspects of subjectivity (e.g., in visual sentiment analysis) or mix subjective components with non-subjective components

(as in adjective-noun pairs) [5]. A problem with the latter is that, in evaluation, these components are mixed and might be hard to separate later on, while the original interest was to focus on subjective parts. Additionally, existing works which use this approach do not include any sophisticated structuring of the subjective components. We also found that there is a clear lack of datasets with more fine-grained or structured subjective aspects annotated.

In order to overcome these shortcomings, we propose a novel dataset (*aspects-DB*, made available to the public¹) and the Focus-Aspect-Value (FAV) model for subjective visual interpretation, disentangling three components of subjectivity: 1) focus: the center of attention, 2) aspect: which dimension to evaluate on, and 3) value: result of this evaluation. Our proposed way of modeling is illustrated in Figure 1. Briefly put, our model works as follows: Given an image and, as context, a noun present in the image (as proxy to describe which part of the image is attended to), we would like to first identify which dimension of evaluation (represented by *aspects*) one is likely to use for describing the noun in the given image, and secondly predict how the noun would be evaluated with respect to these dimensions of evaluation (represented by *aspect values*). Finally, in this paper we analyze several methods for the emerging tasks *aspect prediction* and *aspect-value prediction*, thereby providing an overview of different ways of using context information in this particular case and revealing general open issues.

The rest of the paper is organized as follows. Section 2 surveys related work relevant to this paper. Section 3 introduces the model and the new dataset. Section 4 explains the two tasks that form the core of this method. Section 5 gives a detailed description of the experiments and architectures used. Section 6 provides our insights and findings from the experiments along with the open questions. Section 7 concludes the paper with a summary and future work.

2 RELATED WORK

This section can be broadly divided into three segments: First, the methods which attempt to capture subjectivity. Second, the methods which use adjective-noun pairs for this purpose. Finally, the available attribute detection datasets.

2.1 Detecting subjectivity

There have been many promising approaches which researchers have employed at detecting subjective parts of visual interpretation. While some works focused on attributes to enhance the quality of nouns [6, 12, 22], others focused on understanding the aesthetics [11, 26]. Borth et al. [5, 6] proposed the large scale visual sentiment ontology to detect adjective-noun pairs inside an image. Given an image they propose to find a suitable adjective-noun pair to best describe an image from a set of adjective-noun pairs. Although adjective noun pairs capture the sentiment to an extent, they do not reveal the degree to which this sentiment applies. Moreover, relying on a single adjective-noun pair to describe the whole image would mean only the most prominent noun is focused upon.

Lazaridou et al. [24] propose a cross-modal mapping from a visual semantic space onto a linguistic space in order to automatically annotate images with adjectives. The mapping is performed by a projection function that maps the vector representation of an image

tagged with an object / attribute onto the linguistic representation of the object / attribute word. This mapping function can then be applied to any given image to obtain its linguistic projection. The main advantage, as claimed in their paper, is that of zero-shot learning, i.e., unseen attributes (not present in training) can be predicted. However, in this approach the whole image is mapped onto an adjective without focusing on any particular noun or aspect.

2.2 Detecting adjective-noun combinations

Our work mainly builds on a line of work originating from the Visual Sentiment Ontology [7] proposed by Borth et al., which aims at detecting adjective-noun combinations from images. So far, the best performing method within this direction are cross-residual networks (XResNet) [19], which we include in our experiments and will describe in detail in Section 5.2. For any given image, XResNet outputs scores for adjective-noun combinations as well as scores for all individual adjectives and nouns separately. This means that it separates the more subjective parts of interpretation (represented by the adjectives) from the more objective (represented by the nouns).

There are two major datasets that have been used for training the above-mentioned architectures: The Visual Sentiment Ontology (VSO) [7] and the Multilingual Visual Sentiment Ontology (MVSO) [20]. These datasets have been created from the popular photo-sharing platform Flickr. However, the data in these cases suffers from a clear bias towards the positive attributes / adjectives [21]. In the paper that introduces XResNet [19], some efforts are taken for achieving a better overall balance, but even there, for any given noun the number of associated adjectives is typically very small and the distribution heavily skewed. More importantly, the “feasible” adjectives for a given noun are in most cases not mutually exclusive. At times, adjectives that come with the same noun are even similar in meaning (e.g., “smiling person” and “happy person”), and yet, predicting any adjective that is not identical to the ground-truth adjective is typically considered wrong. For example, “smiling” and “sad” would count as equally bad if the ground truth was “happy”. This makes it harder to interpret performances on these datasets in terms of ability to capture subjective aspects.

2.3 Attribute datasets

There are several popular attribute datasets available for computer vision research. The Visual Genome [23] contains over 100,000 images with fine-grained annotations, including region descriptions, object instances and visual attributes in the order of Millions. However, the attributes in this dataset mostly relate to objective information. Hence, most common attributes are colors like white, blue red, black and despite the large number of total annotations in Visual Genome, we found the number of subjective attribute instances to be too low for our purpose. aPascal and aYahoo [13] are two attribute datasets containing natural object-based images with attribute annotations. Here again, the included attributes correspond to objective features, such as parts of a face like eyes, nose and so on, which deems it inappropriate for analyzing subjective interpretation.

Another attribute dataset is the SUN Attribute Dataset [27], which contains scene attributes of the four categories “functions /

¹Our dataset can be downloaded at <http://madm.dfki.de/downloads>.

affordances” (e.g. “diving”, “climbing”), “materials”, “surface properties” and “spatial envelope”. The former three categories are restricted to objective information, and while there are several subjective attributes (such as “scary” or “stressful”) in the “spatial envelope” category, all of these annotations are describing the scene in a holistic manner. Overall, none of the available datasets is appropriate for focusing on more fine-grained subjective visual interpretation.

3 MODELING AND DATASET

The above-mentioned issues related to adjective-noun-based modeling can be overcome by considering the problem of attribute prediction to focus on the subjective part of visual interpretation: Given an image and an entity (in our case represented by a noun) in the image, estimate the suitability of attributes under consideration of their semantic relations. In this regard we created a dataset with structured attributes for any given noun, thus respecting attribute semantics and enabling a more fine-grained evaluation. Furthermore, our approach allows for zero-shot learning.

3.1 The FAV model for subjective visual interpretation

The work of Borth et al. [7] shows that adjective noun combinations are often visible and reasonably simple to automatically detect in images, presumably because of how they contain both subjective (in the adjective) and objective (in the noun) information.

If we take into account the semantics of adjectives as described by Baroni and Zamparelli in [3], where adjectives are interpreted as modifiers of nouns, we see that visually detecting adjective noun combinations can be understood as a model that combines attention and evaluation, where the noun describes where the viewer is focusing when interpreting the image and the adjective contains the subjective evaluation of this part of the image.

For the adjective, we want to take a step further and acknowledge the fact that adjectives (or “attributes”) for the same noun are often semantically related. In other words, they can be organized along various dimensions of evaluation. Example for such dimensions would be size, age, cuteness or temperature. So instead of considering any non-ground-truth attribute as wrong and thereby largely ignoring semantic relations between attributes, we organize attributes into opposing lists. Arranging in this manner paves way for a more appropriate evaluation, as opposing attributes (mutually exclusive) cannot occur together for the same noun. For example, if we consider the opposing attributes in [“cute”, “adorable”] vs [“scary”, “ugly”], a classification of “cute” or “adorable” of a puppy are semantically similar, but “cute” and “scary” cannot apply to the same puppy. Such a pair of opposing attribute lists reflects a certain dimension of evaluation, which we call “*aspect*” of the noun. Note that an aspect in our case is very similar to the concept of semantic adjective class (used for structuring adjectives in GermaNet [15]), such as *appearance* (“pretty”, “ugly”, ...), *size* (“small”, “big”, “large”, ...) or *age* (“young”, “old”, ...). The only difference is that we group attributes of an aspect into mutually exclusive sets. Some of the specific *aspects* and the adjectives/attributes pertaining to these aspects that are incorporated in our dataset are listed in Table 1 and will be derived in Section 3.2.

In summary, we separate three potential sources of subjectivity in the FAV model:

- (1) *Focus*: Given a single image, there are typically different components one can pay attention to. For this paper we will assume that this place of focus can be captured by a noun. Note that nouns can relate to an entity in the image (such as “dog” or “dude”), but also refer to the whole scene (as in “place”) or the picture itself (“shot”).
- (2) *Aspect*: Once the focus has been determined, there are several potential dimensions for evaluation. For example, people in the image can be evaluated with respect to their physical size, age, level of activity and so on. In our dataset, selecting an aspect for evaluation is essentially about choosing a set of semantically related attributes.
- (3) *Value*: For this paper, we chose all aspects to be represented by two mutually exclusive sets of adjectives, such that evaluating each aspect amounts to a binary decision problem. For example, physical size would have adjectives like “small”, “tiny”, “short” on one side and “tall”, “big”, “huge” on the other. Picking a certain value then means to select a set of attributes that are appropriate to be used as an attribute for the given noun.

The following points summarize the key features of this method of modeling:

- Three different sources of subjectivity are disentangled. This brings about the possibility to evaluate these components separately, and helps to make results easier to understand. Readers familiar with NLP literature might recognize the analogy to opinion mining, a task where topic, aspect and sentiment are extracted from text in order to explain prevalent opinions (see e.g., the survey of Liu and Zhang [25]).
- Semantic relations between attributes are respected. In particular, by detecting aspect values instead of individual attributes, we treat attributes of the same value as being synonymous for the given aspect. We thereby avoid to consider any attribute as wrong if it means the same but is merely phrased differently, as it is for example done when using adjective noun combinations or single attributes as independent class labels.
- This modeling leads to a more sensible way of 0-shot learning for attribute detection, i.e., predicting subjective attributes for nouns for which they were not available during training time. We will explore this direction below.

3.2 Compiling the dataset

To overcome the shortcomings of the existing datasets mentioned above, and to have a fair evaluation for experiments, we decided to create a new dataset called *aspects-DB* for subjective visual interpretation, following the FAV model. We will now describe the steps we took for building the dataset.

Our dataset is build from Google’s Conceptual Caption Dataset [29], which contains over 3 Million images together with natural-language captions. First, we ran a POS tagger (using NLTK [1]) over all these captions. From these POS-tagged captions, we compiled a list of all adjectives and nouns which appear in adjective-noun combinations (we selected the adjective-noun combinations which

No.	Name	Values		#Images	#Nouns
1	color	COLORFUL ("blue", "turquoise", "green", "colorful", "red", "purple", "coloured", "colored", "golden", "yellow", "silver", "orange", ...)	COLORLESS ("white", "black", "gray", "grey", "bland")	19914 (6728 vs 13186)	46
2	age	YOUNG ("modern", "new", "young", "trendy", "youthful", "teenage", "teen", "contemporary", "current", "recent")	OLD ("old", "historic", "colonial", "medieval", "ancient", "historical", "traditional", "elderly", "senior", "aged", "vintage", ...)	19793 (11169 vs 8624)	70
3	size	SMALL ("small", "tiny", "little", "miniature")	BIG ("large", "giant", "big", "huge", "massive", "major", "grand", "enormous", "oversized", "astronomical")	19177 (12918 vs 6259)	78
4	sun	SUNNY ("sunny", "bright", "clear")	CLOUDY ("cloudy", "rainy", "misty")	14321 (10641 vs 3680)	21
5	rareness	UNUSUAL ("unique", "ornamental", "creative", "different", "oriental", "unusual", "exotic", "popular", "stylish", "magical", ...)	ORDINARY ("local", "daily", "typical", "generic", "general", "regular", "familiar", "casual", "usual", "similar", "normal", "natural", ...)	13299 (6804 vs 6495)	94

Table 1: Five most common aspects (out of 19) in the *aspects-DB* dataset. We only list attributes that are included in any adjective-noun combination in the dataset. Numbers in parentheses indicate how many images the dataset contains for the two possible values. The remaining aspects in *aspects-DB* are (from most common to least common): AERIAL vs PANORAMIC, FIRST vs LAST, SMILING vs SAD, FRONT vs BACK, INTERIOR vs EXTERIOR, BRIGHT vs DARK, RURAL vs URBAN, HOT vs COLD, TINY vs TALL, GOOD vs BAD, BUSY vs LAZY, PRIVATE vs PUBLIC, OPEN vs CLOSED, WESTERN vs EASTERN.

have appeared at least 200 times in the dataset). Our underlying assumptions were that whenever we find such an adjective-noun combination inside a caption (e.g., "cute puppy"), it is very likely that the adjective describes a property of the noun (or "attribute") and that the noun is visible in the image.

Next, we manually organized all resulting adjectives (that are potentially visible) into *aspects*. This gave us an initial list of aspects with associated attributes (represented by adjectives) grouped into mutually exclusive sets.

We now collected all images from the Conceptual Caption Dataset, which have an adjective-noun combination in their caption where the adjective is included in any one of our aspects. This gave us an initial dataset of over 400,000 images together with associated adjective-noun combination, aspect, and aspect value for each image. We then manually went through the list of remaining nouns, and excluded some words ("retro", "beautiful", "news", "cloudy") which were falsely labeled as noun by the POS tagger, or not clearly visible in images. Finally, we iteratively removed data until all the following criteria were satisfied:

- For each noun-aspect combination, there are at least 10 images for each value.
- For each aspect, there are at least 500 images for each value (across all nouns).
- For each noun, there are at least 50 images for each value (across all aspects).
- For each aspect, there are at most 20,000 images in total. (For aspects with more images, we did a simple down-sampling to reduce the number.)

The final *aspects-DB* dataset contains 155,539 images in total and features 143 nouns for 19 aspects. A list with the 5 most common aspects can be found in Table 1. Since the ground truth was obtained by adjective-noun pairs, we keep the adjective part in our dataset as extra information, so for each image, *aspects-DB* includes a noun, an aspect, the value of this aspect and the original adjective the noun was combined with in the caption. Table 2 shows a few examples of ground truth information for two particular aspect-noun combinations. The dataset is available to the public and can be downloaded at <http://madm.dfki.de/downloads>.

We would like to emphasize that the ground truth in *aspects-DB* is meant to capture general tendencies in subjective interpretation (where we use tags as proxy). These tendencies must to some extent be corpus / domain specific and on item-level we cannot expect perfect performance. This means that the task is not to detect objectively correct labels as in many common image classification datasets, but to model general biases such as "for this image of a sleepy puppy and noun *dog*, people would typically interpret the image with respect to aspect *age*. Aspects *age*, *activity*, *evaluation* would likely be rated as having values *young*, *sleepy*, *good* respectively".

4 TASKS

4.1 Aspect prediction

In the first task, an image and a noun are given and the task is to predict which one of the aspects in our dataset (see Table 1) a subjective interpretation would most likely focus on. For example,

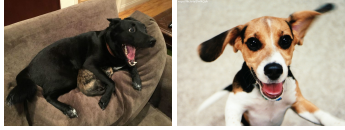
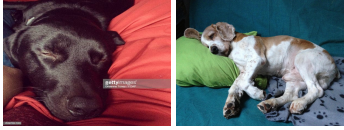


Aspect	Noun	Value	
activity	dog	ACTIVE	LAZY
			
size	tree	SMALL	BIG
			

Table 2: Examples of ground truth data in the proposed *aspects-DB* dataset. Each row represents an aspect for an example noun, and contains sample images which corresponds to the two possible values. All images are part of Google’s Conceptual Caption Dataset [29].

given an image with a puppy together with the noun “dog”, a likely aspect from our list would typically be *age*. This problem is modeled as multi-class classification task, where for each given image and noun, only a single aspect is considered to be correct. We evaluate in terms of overall prediction accuracy.

4.2 Aspect Value Prediction

Aspect Value Prediction is about deciding which value applies to a given noun for a given aspect in the context of the input image. Coming back to the previous puppy example of Section 4.1, the true value for aspect *age* would be YOUNG when given an image of a puppy with the noun context “dog”. For training and evaluation we only consider one aspect at a time, hence this problem can be seen as binary classification task.

In *0-shot* value prediction, evaluation is done on noun-aspect combinations that were not available during training time. (Ground truth data for other noun-aspect combinations with the same noun but different aspects or same aspect but different nouns is assumed to be available for training.) For calculating overall accuracy for value prediction, we compute accuracy over all test set items.

4.3 Dataset split

We use two different dataset splits, the *standard* split and the *0-shot* split.

- For the *standard* split, all available data is for each value randomly split into 60% training, 20% development and 20% test data. This implies that aspect and value priors are identical for training, development and test set. We use the standard split for experiments on aspect prediction and aspect value prediction.
- For the *0-shot* split, we randomly split noun-aspect combinations, using 60% of the combinations for training, 20% for development and 20% for testing. We use this dataset split for experiments on aspect value prediction. It should be noted that 0-shot learning on aspect prediction cannot be done in the same way (unless the noun is left out completely for

training): If we remove individual noun-aspect combinations and train a model on the remaining ones, the model generally learns that for any noun the excluded aspects are not feasible. This points to another problem in the adjective-noun way of modeling, where aspect and aspect value are both blended into adjective information.

5 METHODS

In this section we explain the methods we compare in our experiments (Section 6), where they are evaluated on both tasks described in the previous section. For all models, except the XResNet variants, visual features are extracted from the image by a ResNet-50 network [17], which was trained on ImageNet [10] and kept unchanged.

5.1 Logistic regression

As baselines, we deploy two models based on logistic regression which take visual features from the inception network as the only input:

- The *noun-agnostic* version does not consider noun information at any point. Aspect prediction is modeled as classification task with multiple classes. So for predicting the most likely aspect given an image and noun, a single logistic regression model is trained to output the corresponding class from the visual features, irrespective of the noun. Aspect value prediction is modeled as separate binary classification problems, i.e., for each aspect, one logistic regression model is trained to detect the value of the respective aspect from the image vector, again not taking the noun into account.
- In the *noun-specific* variant, separate models are trained for distinct nouns. For each individual noun, we then follow the same approach as described in the previous point. This means that for each noun we have one model predicting the most likely aspect, and for each noun-aspect combination we have one model for aspect value prediction. We explore this possibility as a simple way to take the noun context into account.

In both cases we use the scikit-learn [28] implementation for training and inference.

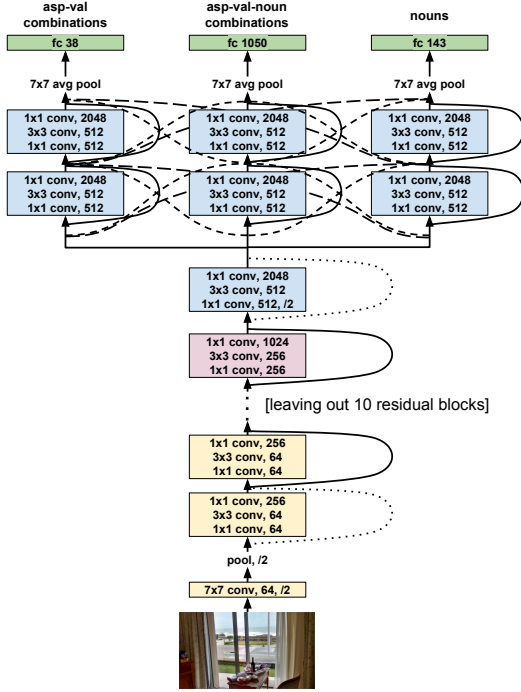


Figure 2: XResNet architecture (adapted from [19]). Solid shortcuts indicate identity, dotted connections indicate 1×1 projections, and dashed shortcuts indicate cross-residual weighted connections.

5.2 Cross-residual networks

Cross-residual network, or short XResNet, refers to an architecture which was introduced by Jou and Chang [19] for adjective-noun pair detection and is based on the well-known residual network (ResNet) architecture [17]. Figure 2 shows the structure of the XResNet architecture we used. The main difference of XResNet as compared to ResNet is that the network branches out at the end into three distinct heads, where these branches remain closely connected to each other via so-called cross-residual connections. The standard XResNet architecture has 50 layers and finally branches out to predict adjectives, nouns and adjective-noun pairs respectively.

We adapt XResNet to our tasks, by replacing these original output branches by three branches specific to our tasks. Instead of adjectives, one branch outputs scores for all combinations of aspect and aspect value. Instead of adjective-noun pairs, scores for all combinations of aspect, aspect value and noun are predicted. The noun branch remains unchanged. This leaves us two possibilities for evaluation:

- The final decision can be made based on the aspect-value branch (*asp-val*): For aspect prediction, we use the aspect of the aspect-value combination with the highest score as

output. In case of value prediction for a given aspect, the value of the aspect-value combinations with the given aspect and highest score is considered. Note that in this version the noun context is ignored.

- The other version is based on the aspect-value-noun branch (*asp-val-noun*): First, all combinations with irrelevant nouns are removed. The rest is done completely analogous to the *asp-val* case.

5.3 Concatenation + MLP

The concatenation model is a straightforward application of information fusion, where a one-hot encoding of the noun is appended to the image embedding obtained from the inception network. This concatenated vector is then used as the input to a multi-layer perceptron (MLP) with one hidden layer. The MLP has two output branches, one for aspect prediction and one for detecting aspect value. More precisely, the hidden activation h is computed as

$$h(x, n) = \tanh \left([W_1 | W_2] \cdot \begin{bmatrix} x \\ n \end{bmatrix} + b \right) = \tanh (W_1 x + W_2 n + b)$$

where $\begin{bmatrix} x \\ n \end{bmatrix}$ stands for the concatenation of x and n , b is a bias vector, and W_1, W_2 are weight matrices of suitable shapes. The dimension of h is a hyper-parameter and is referred to as number of hidden units. The model then estimates aspect likelihoods as

$$\text{softmax}(W_a \cdot h(x, n) + b_a)$$

and aspect values as

$$\tanh(W_p \cdot h(x, n) + b_p),$$

where b_a, b_p are bias vectors, and W_a, W_p weight matrices of suitable shapes such that the output dimension for both branches is equal to the number of aspects. Note, however, that for value prediction during training and testing we only consider the output of the unit corresponding to the aspect which is processed at the time.

5.4 Tensor Fusion

Instead of merely concatenating image features and context we consider a slightly more sophisticated way of conditioning on the context, where higher-order interactions between input and context information are described by a weight tensor. Similar ways of using context information have been used in several publications in the field of natural language processing (for example [2, 14, 16]). In computer vision, a related approach for information merging can be found in the MUTAN model [4] for question answering, where question and image embeddings are merged under the use of Tucker decomposition.

The Tensor Fusion approach is illustrated in Figure 3. The core part is the *Tensor Fusion layer*, which can be understood as part of a neural network that combines the noun-agnostic and noun-specific logistic regression models: For each noun $i = 1, \dots, 13$, there is a weight matrix W_i and a bias term b_i . In addition, the layer uses a weight matrix W_0 and a bias term b_0 that are independent of the noun context. Given as input the image embedding x and the i -th noun, the output of the Tensor Fusion layer is then computed as

$$(W_0 + W_i) \cdot x + b_0 + B_i.$$

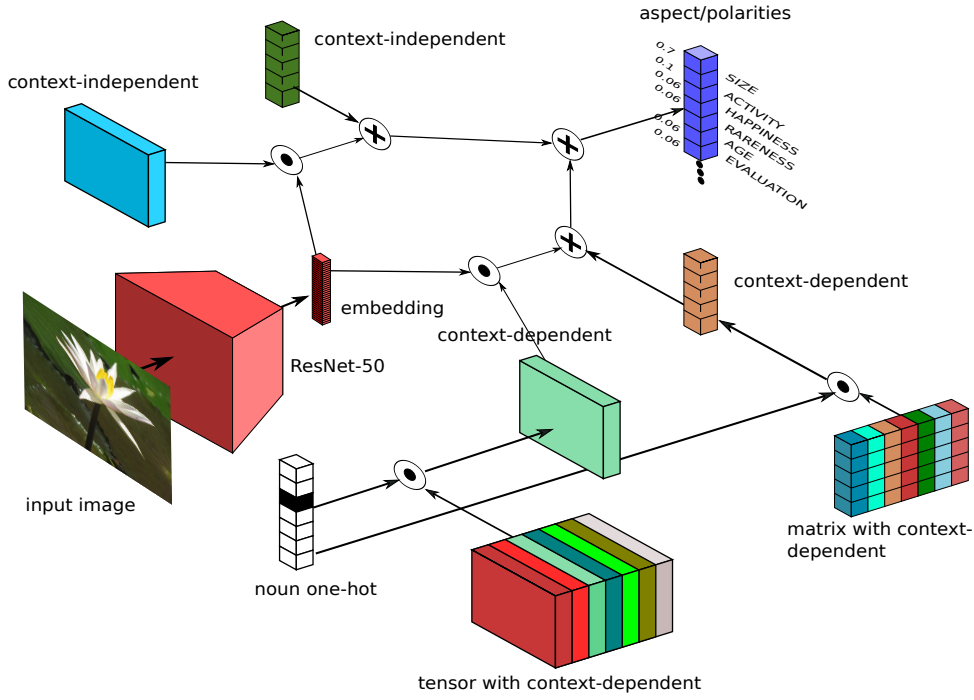


Figure 3: An overview of the Tensor Fusion model. Given as input a one-hot encoded noun and an image, the Tensor Fusion model embeds the image with a pre-trained ResNet network. This image embedding vector is then processed with the Tensor Fusion layer which consists of two linear layers, a context-independent and a context-dependent one, followed by element-wise additive fusion of their outputs. For the context-dependent path, the Tensor Fusion layer keeps a tensor with context-dependent weights and a matrix with context-dependent biases, which are multiplied by the noun vector to obtain weights and bias. (Since the noun is one-hot encoded, this multiplication amounts to a selection operation.) We use two separate Tensor Fusion models for our experiments, one for aspect prediction with aspect likelihoods and one for aspect value prediction with aspect values as output.

We now represent nouns as one-hot vectors $n \in \mathbb{R}^{143}$ and put together all noun weight matrices W_i into a third-order Tensor $W \in \mathbb{R}^{143 \times 1000 \times 19}$ and all noun biases b_i into a bias matrix $B \in \mathbb{R}^{143 \times 19}$. The final layer function $T(x, n)$ can be formulated by using a multiplication operation between the noun context and the weight tensor to obtain the weight matrix for the given noun:

$$\begin{aligned} T(x, n) &= \left(W_0 + \sum_{i=1}^{143} W_i \cdot n_i \right) \cdot x + b_0 + \sum_{i=1}^{143} B_i \cdot n_i \\ &= (W_0 + W \circ n) \cdot x + b_0 + B \cdot n, \end{aligned}$$

where $W \circ n := \sum_{i=1}^{143} W_i \cdot n_i$.

We deploy separate Tensor Fusion models for the two tasks of aspect prediction and aspect value prediction.

5.5 Linear fusion

Using the same notation for variables as above, we define the linear fusion layer as having the output

$$\text{linear}(x, n) = W_0 \cdot x + b_0 + B \cdot n.$$

This enables another view of the Tensor Fusion layer, namely to interpret it as linear fusion plus a term for capturing higher-order interactions:

$$\begin{aligned} T(x, n) &= W_0 \cdot x + b_0 + B \cdot n + (W \circ n) \cdot x \\ &= \text{linear}(x, n) + (W \circ n) \cdot x \end{aligned}$$

Hence, we include linear fusion into our experiments in order to single out the role of the higher-order interaction term, which makes the majority of trainable parameters for Tensor Fusion.

6 RESULTS

For both tasks described in Section 4, we ran experiments with all conditioning methods explained in Section 5. All results are listed in Table 3. Hyper-parameters (learning rate, regularization weight, and number of hidden units for the concatenation method) were optimized based on performances on training and development data (see Section 4.3). We report performances on the test data.

Approach		Aspect accuracy (standard split)	Value accuracy	
Model	Variant		standard	0-shot
logistic regression	noun-agnostic	65.67%	78.97%	60.45%
	noun-specific	67.07%	84.79%	-
XResNet	asp-val	45.64%	80.36%	63.71%
	asp-val-noun	70.30%	85.34%	-
concatenation + MLP	100 hidden units	50.89%	79.36%	68.43%
	5000 hidden units	53.23%	80.11%	61.24%
linear fusion	N/A	62.15%	80.34%	68.41%
Tensor Fusion	N/A	69.46%	86.34%	69.04%

Table 3: Aspect prediction and aspect value prediction performances of all models. All methods except the XResNet models use a pre-trained ResNet to embed the image. Please refer to Section 5 for details on the individual models. Note that not all models are applicable to the 0-shot learning task.

6.1 Aspect prediction

For aspect prediction, XResNet turned out to be the best performing method (70.30% for *asp-val-noun*), closely followed by Tensor Fusion (69.46%). Both of these models outperform the logistic regression baselines (67.07% for noun-specific and 65.67% for noun-agnostic). The linear fusion model performs worse than the noun-agnostic logistic regression baseline (62.15%). This is unexpected because this model essentially computes the same as noun-agnostic logistic regression plus a linear part coming from the noun. We assume that this effect is due to differences in training and implementation, which was done with the sklearn implementation for logistic regression and using an own neural network implementation in tensorflow for the linear fusion model. The gap between performances of the linear fusion to the Tensor Fusion model shows clearly that incorporating higher-order interactions between image features and noun context is beneficial to the task at hand.

Interestingly, concatenation yields very poor performances for aspect prediction. An accuracy of 10% worse than the linear fusion model suggests that the concatenation models were not able to properly make use of the noun information. Looking at the training behavior (not shown here), we also found that for the same aspect prediction training loss of 1.26, linear fusion achieved accuracies around 60% while accuracies of concatenation was around 50%, so concatenation seems much more prone to overfitting.

6.2 Aspect value prediction

With both the standard dataset split and the 0-shot split, Tensor Fusion gave the best results for aspect value prediction (86.34% for standard, 69.04% for 0-shot). Using the standard split, XResNet was able to achieve comparable performance when using the *asp-val-noun* output branch (85.34%). For detecting values of unseen noun-aspect combinations (0-shot column), however, XResNet has to rely on the *asp-val* output, and clearly falls behind Tensor Fusion, linear fusion and one of the concatenation models.

Noun-specific logistic regression is almost on par with XResNet for the standard task (84.79% vs 85.34%), but cannot be applied to 0-shot learning, where the noun-agnostic logistic regression model lead to the lowest overall accuracy (60.45%).

Linear fusion, concatenation, the *asp-val* version of XResNet and noun-agnostic logistic regression all yield comparable performances (between 78.97% and 80.36%), around 6% lower than the top performing methods. Surprisingly, the corresponding 0-shot results of these models show much greater variation. In particular, concatenation with 100 hidden units gives the second-highest overall score for the 0-shot experiment (68.43%), but the second-lowest one for the standard task (79.36%).

7 CONCLUSION

We introduced a new approach for capturing subjectivity prevalent in images. To overcome several challenges, including the heavy bias towards positive tags / titles in social media, and to make it possible to separately evaluate different parts of subjective visual interpretation, we compiled a new dataset based on Google’s recently released Conceptual Captions Dataset [29]. We ran our experiments on the new dataset and reported results with different architectures. It was also shown that with the new modeling, it is possible to perform *0-shot* learning to predict unseen noun-attribute combinations. Given the prevalence of simple concatenation for combining information in deep learning approaches, we find it interesting that Tensor Fusion performed better across experiments.

Our results raise some fundamental questions, which we want to investigate in the future: a) How can context be modeled optimally? Often researchers use concatenation as default choice and focus on data or hyper-parameters for improvement without changing this part of the architecture, but our results showed a decrease in performance when using the concatenation method; b) Which properties of the tasks make some methods (like concatenation) fail in one but be among the best performing methods in another?

Furthermore, we plan to explore other ways of conditioning on context, and adapt our approach to applications such as personalized tag prediction and affective image captioning, where biases at different stages of subjective visual interpretation according to the FAV model can be made dependent on a user context to mimic the subjective interpretation of the given user.

ACKNOWLEDGMENTS

This work was supported by the BMBF project DeFuseNN (Grant 01IW17002) and the NVIDIA AI Lab (NVAIL) program.

REFERENCES

- [1] 2018. Python-NLTK Documentation. <https://www.nltk.org/book/ch05.html>.
- [2] David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed Representations of Geographically Situated Language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 828–834. <https://doi.org/10.3115/v1/P14-2134>
- [3] Marco Baroni and Roberto Zamparelli. 2010. Nouns Are Vectors, Adjectives Are Matrices: Representing Adjective-noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1183–1193. <http://dl.acm.org/citation.cfm?id=1870658.1870773>
- [4] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, Vol. 3.
- [5] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 459–460.
- [6] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 223–232.
- [7] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 223–232. <https://doi.org/10.1145/2502081.2502282>
- [8] Heinrich H Bulthoff. 1996. Bayesian decision theory and psychophysics. *Perception as Bayesian inference* 123 (1996).
- [9] Claus-Christian Carbon. 2011. Cognitive mechanisms for explaining dynamics of aesthetic appreciation. *i-Perception* 2, 7 (2011), 708–719.
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255.
- [11] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1657–1664.
- [12] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 1778–1785.
- [13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. 2009. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 1778–1785. <https://doi.org/10.1109/CVPR.2009.5206772>
- [14] Emiliano Guevara. 2010. A Regression Model of Adjective-noun Compositionality in Distributional Semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics (GEMS '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 33–37. <http://dl.acm.org/citation.cfm?id=1870516.1870521>
- [15] Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. 9–15.
- [16] Matthias Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, 54–64. <http://aclweb.org/anthology/E17-1006>
- [17] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [18] Guido Hesselmann, Christian A Kell, and Andreas Kleinschmidt. 2008. Ongoing activity fluctuations in hMT+ bias the perception of coherent visual motion. *Journal of Neuroscience* 28, 53 (2008), 14481–14485.
- [19] Brendan Jou and Shih-Fu Chang. 2016. Deep Cross Residual Learning for Multi-task Visual Recognition. In *ACM Multimedia*.
- [20] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 159–168.
- [21] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. 2015. Real-time analysis and visualization of the YFCC100M dataset. In *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*. ACM, 25–30.
- [22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.
- [23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. <https://arxiv.org/abs/1602.07332>
- [24] Angeliki Lazaridou, Georgiana Dinu, Adam Liska, and Marco Baroni. 2015. From visual attributes to adjectives through decompositional distributional semantics. *TACL* 3 (2015), 183–196.
- [25] Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, 415–463.
- [26] Anush K Moorthy, Pere Obrador, and Nuria Oliver. 2010. Towards computational models of the visual aesthetic appeal of consumer videos. In *European Conference on Computer Vision*. Springer, 1–14.
- [27] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *International Journal of Computer Vision* 108, 1-2 (2014), 59–81.
- [28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypersynthesized, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of ACL*.
- [30] Marie L Smith, Frédéric Gosselin, and Philippe G Schyns. 2012. Measuring internal representations from behavioral and brain data. *Current Biology* 22, 3 (2012), 191–196.

Image Captioning in the Wild: How People Caption Images on Flickr

Philipp Blandfort^{*†‡}, Tushar Karayil^{*†}, Damian Borth[†], Andreas Dengel^{†‡}

[†] German Institute for Artificial Intelligence, Kaiserslautern, Germany.

[‡] University of Kaiserslautern, Kaiserslautern, Germany.

[first_name].[last_name]@dfki.de

ABSTRACT

Automatic image captioning is a well-known problem in the field of artificial intelligence. To solve this problem efficiently, it is also required to understand how people caption images naturally (when not instructed by a set of rules, which tell them to do so in a certain way). This dimension of the problem is rarely discussed. To understand this aspect, we performed a crowdsourcing study on specific subsets of the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) where annotators evaluate captions with respect to subjectivity, visibility, appeal and intent. We use the resulting data to systematically characterize the variations in image captions that appear “in the wild”. We publish our findings here along with the annotated dataset.

KEYWORDS

YFCC100M; Flickr; image captioning; subjectivity; intent; sentiment

1 INTRODUCTION

A caption for an image is a short piece of text, usually a one-liner, provided by the user and describes his/her interpretation of the image. This interpretation can vary from being very objective to being very subjective or even poetic. But subjectivity is only one out of several possible dimensions for analyzing variations in image captioning.

In order to enable computers to more effectively interact with humans in multimodal environments such as social media platforms, a better understanding of these and other variations could be highly beneficial. Moreover, with the surge of end-to-end deep learning methods, finding representative datasets for the task at hand has become even more important. Here, statistics of captioning in the wild are potentially useful for obtaining such data and gaining awareness of different forms of prevalent biases.

So the question is how does “captioning in the wild” look like and how can the variety in the captions be captured formally? To this end, we performed a crowdsourcing study for making sense of

* Equal contribution from both authors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MUSA2’17, October 27, 2017, Mountain View, CA, USA.
© 2017 ACM. ISBN 978-1-4503-5509-4/17/10...\$15.00
DOI: <https://doi.org/10.1145/3132515.3132522>

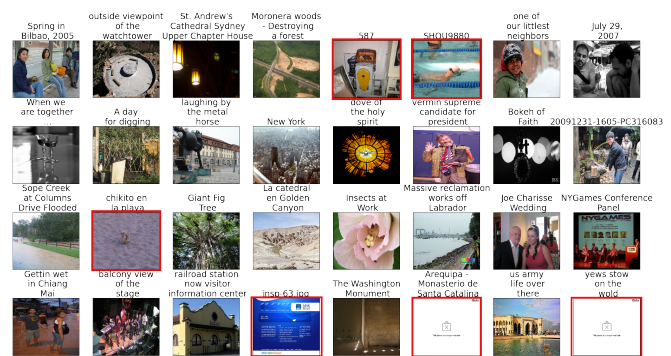


Figure 1: Illustration of randomly sampled image caption pairs from the YFCC100M dataset to show the variation in captions. The captions found here vary from something as simple as a number (e.g. “587”) to complex subjective interpretation (e.g. “when we were together”). The empty white images on the bottom row constitute one class of noise images present in the dataset wherein the user has removed the corresponding image. The images marked with a red boundary represent some examples of noise in the dataset for this task.²

this diversity. For this study we have generated 3 subsets of Flickr¹ images with English captions and asked people across the English speaking world to annotate these image/captions pairs to evaluate them based on measures like subjectivity, visibility of information and intent.

To generate the study dataset, the authors have used YFCC100M [15]. YFCC100M was introduced in 2014 and is, to date, the largest collection of publicly available multimedia data. It contains 99.2 million photos and 0.8 million videos. All of this content has been taken from Flickr between the years 2004-2014. Having sampled from Flickr, which is a common platform used for photo-sharing, the YFCC100M contains a diverse collection of real world videos, pictures, tags, captions and descriptions. All this data is user-generated, which makes it different from other datasets like Microsoft-Common Objects in Context [9] or Flickr30k [12] which were built with the help of paid annotators. Figure 1 shows a few examples of the kind of captions found across YFCC100M.

¹<https://www.flickr.com/>

²Images are licensed under CC. Corresponding authors can be found at <https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67>.

The diversity in YFCC100M also presents its own challenges [11, 15]. For example, the presence of generic titles like camera generated ones or non-English languages [8] also add to the high quantity of noise in the dataset. Therefore, a filter to weed out the noise is essential for this task. The design of the filter should be such that it not only filters out the noise, but also retains meaningful variations in the captions.

In this regard the following are the main contributions of this paper:

- We introduce a user annotated dataset for studying caption variations in the wild. We make this dataset publicly available to aid further research.³
- We perform a quantitative analysis on this dataset with the aim of characterizing how people naturally caption images.
- We show how simple filtering techniques can be used to remove useless titles from YFCC100M.

The rest of the paper is organized as follows. Section 2 surveys related work relevant to this paper. Section 3 explains the data source and the different filters used for the crowdsourcing experiment. Section 4 gives a detailed description of the experiment settings including the steps taken to assure quality and the annotator backgrounds. Section 5 provides our insights and findings from the collected annotations. Section 6 concludes the paper and charts out the future direction.

2 RELATED WORK

Although there have been numerous works in automatic caption generation [3, 5, 7, 17, 18], to the best knowledge of the authors, there has only been little work on systematically analyzing caption variations.

There have been some studies which have tried to gain insights into the language and other statistics specifically for the YFCC100M. The authors of [8] have performed an analysis on the distributions of languages and geo-locations in the YFCC100M. They reported various characteristic features like spoken language in titles, descriptions, tags etc. Another study performed by [6] provides a general overview of the YFCC100M and also provides statistics of adjectives and nouns present in the dataset. Furthermore, works by [19, 20] built the “The Image-Emotion-Social-Net Dataset” which includes a large number of image comments from Flickr annotated by emotions that they are supposed to convey. However, the focus of their work is more on providing data for emotion prediction models and not on describing captioning behavior.

Apart from these, most studies on Flickr or YFCC100M have been primarily focused on tags. The authors of [1, 2, 13] have for example studied image tags to understand the semantics of data for index generation. Studies in [16] and [14] have analyzed Flickr language models for geo-location.

In contrast to the previously listed works, this paper focuses on investigating captioning styles across Flickr and YFCC100M dataset. It provides unique insights into the captioning styles present in the wild which are of high relevance for the automatic image captioning and language analysis community.

³The dataset can be found on <http://madm.dfki.de/downloads>.

3 DATA AND FILTERING

Flickr is a social platform, where users can share images and videos. This multimedia content is further enriched by a set of optionally annotated metadata in the form of tags, titles, descriptions provided by the user. The YFCC100M⁴ consists of almost 100 million such images compiled from the Creative Commons multimedia content of Flickr. Each of these images also come with the following metadata: author, date, geolocation, title, description, tags and EXIF information.

A variety of languages can be found in the metadata of the YFCC100M (see [8]). For this reason we also need to filter for language in addition to other noise in the caption in the form of: a) generic titles, b) short titles (depending on purpose), c) merely numeric titles (e.g. only give date of the image)

For this reason we first decided to use the MM Commons Yahoo-Flickr Grand Challenge⁵ filter which performs the following steps: (1) All punctuation characters are removed. (2) All letters are converted to lower case. (3) If the caption has less than 5 words it is removed. (4) If any of the words in the caption does not appear in an English dictionary, the caption is dropped.

This filter, of course, was designed for the specific purpose of the grand challenge and may not be the right choice for other tasks. To address this bottleneck, we also designed an own simple filter to overcome the shortcomings of the grand challenge filter. We removed the noise using the following criteria (checked in this order):

- ‘empty’: The caption is an empty string.
- ‘generic’: The caption is generic (e.g. IMG_123). This is checked using a regular expression.
- ‘short’: There are less than min_words words in the caption. (For this study we used a threshold of 2.)
- ‘non-en’: The caption is not English. To compute this we remove most punctuation characters and for the resulting string check whether the fraction of words that are contained in the enchant dictionary is greater than or equal to a given threshold. (Set to 0.5 for this study.)
- ‘numeric’: The main part of the caption is made of numeric expressions, such as single numbers or dates.
- ‘valid’: All captions that don’t fall into any of the previous categories are considered to be useful and pass the filter.

Note that for our filter the original caption is kept (even though in intermediate steps it might be modified).

4 EXPERIMENTAL SETTING

4.1 The Task

We performed a crowdsourcing experiment on the annotation platform CrowdFlower⁶. The task involved annotation of three subsets of the YFCC100M, each containing 1000 image/caption pairs, sampled from the following selections:

- The unfiltered YFCC100M
- The MM Commons Yahoo-Flickr Grand Challenge dataset⁷
- YFCC100M filtered by our filter outlined above

⁴<https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67>

⁵<https://multimediacommons.wordpress.com/tag-caption-prediction-challenge/>

⁶<https://www.crowdfLOWER.com/>

⁷<https://multimediacommons.wordpress.com/tag-caption-prediction-challenge/>

These subsets were combined and shuffled to create a dataset of 3000 Image/Caption pairs for annotation.

We collected annotations for the following fields:

- Image error: Some images become unavailable and answering the remaining questions is not very useful for these cases.
- Preference: Annotators are asked how much they personally like the given image title on a scale from 1 (“not at all”) to 5 (“very much”).
- No English: Having captions of different languages annotated is hardly possible within a single crowdsourcing study. In this study we focus on English captions.
- Subjectivity⁸: The question “How subjective is the title?” is answered by rating the image title on a scale from 1 (for “purely objective”) to 5 (for “purely subjective”). High subjectivity indicates that there are many other options for captioning the given image.
- Visibility⁹: The question “How much of the information given in the title can be seen in the image?” is answered by rating the image title on a scale from 1 (for “nothing”) to 5 (for “all of it”). Information from the title (e.g. people, objects, famous landmarks etc.) is considered to be “visible” if it can be directly identified in the image. Among other things, visibility is important for computational feasibility.
- Understanding: For being able to filter out captions that were not understood by our annotators, we added a field for specifying a lack of understanding.
- Intent: The last task is to specify in which situations one would most likely use such a title for the given image. This is done with respect to the intent categories “to entertain someone” (e.g. in a humorous, witty, artistic or poetic way), “to provoke someone” (e.g. insult someone, tease someone or draw public attention), “to report *factual* information”, “to express emotions or an attitude” and “other”. This information is particularly useful when filtering for a specific image captioning purpose.

A screenshot of the interface for the complete task is shown in Figure 4.

4.2 Quality Assurance

The experiment ran for a span of 6 months and had a total of 298 annotators participating at various stages. Each annotator was allowed to annotate a maximum of 100 captions.

As the task involved fairly difficult questions like evaluating the subjectivity present in the captions, the annotators were given detailed instructions on the task, with not only an articulate explanation of each field but also a significant example of an image/caption pair for each possible option for the fields “subjectivity”, “visibility” and “intent”.

⁸By “subjective” we mean based on or influenced by personal feelings, tastes, or opinions, as opposed to “objective” which relates to facts such as names, dates or other factual information.

⁹Famous landmarks (e.g. London Bridge, Statue of Liberty, Eiffel Tower etc.) can be considered as visible since they can be identified by most of the people, whereas for example pet names are non-visible information since only a few people are able to identify the pets.

Furthermore, annotators were first asked to tackle a set of test questions designed by the authors.¹⁰ Only if they scored satisfactorily, were they allowed to proceed to the actual task. To ensure that a sufficient annotation quality was maintained throughout the annotation task, test items were intermixed with the actual items and annotators who went above a threshold of error rate on the test items were dropped. The annotators were also required to be from English speaking countries. Each item was evaluated by 5 different annotators.

5 ANALYSIS

5.1 General Observations

After removing noise from the data, there is still a huge variety in the remaining captions. This is due to several reasons.

First of all, captioning itself can be understood in several ways and depending on this understanding the caption will be shaped accordingly. For example, we found that many captions are titles in a rather “classical sense”, i.e. they consist of only a few words which could be understood as a distinguishing name for the image. Names of places or people are often used to label the image in this way, but more artistic names are not uncommon as well. Similarly, image titles can be used to merely summarize what is in the image. This is essentially a translation process from vision to language and captions of this type have been the focus for the vast majority of papers in the field of automatic image captioning. More generally, image captions can be considered to be a part of a multimodal message consisting of text and image. Here, captioning is part of a communication process. So, in this setting, captioning an image would mean to find a suitable phrase or short sentence that together with the image forms a single message and conveys some desired meaning. This meaning could be a subjective interpretation of the image but captions that provide context information (such as names of places or what happened just before the image was taken) or guide the viewer towards a certain interpretation of the image (e.g. by mentioning certain aspects of the image one would not typically see at first glance) also fit nicely into this way of modeling.

Another major source for variation is the interpretation of the image. Roughly, this interpretation can be modeled as a two stage process where individual differences can be observed at each of the two stages:

- (1) Topic/focus: The caption can relate to parts of the image, the image as a whole, or to some external context. So the first step is to focus on some aspect of the image or the context.
- (2) Evaluation: This aspects of the image is then interpreted which gives the final information that will appear in the caption. This evaluation can happen in a subjective way (e.g. liking or disliking parts of the image or the image as a whole) or be of a more objective sort (e.g. remembering when the house in the image was built and putting this into the caption).

Finally, individual language styles, i.e. how the authors formulate the meaning they want to convey, are one more source for variation.

¹⁰For the test questions we used items from the same annotation task, where for each item we specified answers which are acceptable.

	image error	no English in title	title not understood	clean captions
unfiltered YFCC100M	4.6% (46 out of 1000)	51.2% (488 out of 954)	10.52% (49 out of 466)	417
our filter	2.4% (24 out of 1000)	4% (39 out of 976)	4.91% (46 out of 937)	891
grand challenge	3.5% (35 out of 1000)	0% (0 out of 965)	3.53% (35 out of 965)	930

Table 1: Noise distribution and number of useful captions for all selections, based on majority votes. Note that captions with image errors have been removed prior to calculating percentages for the “no English in title” column. Similarly, for calculating the “title not understood” percentages captions with either an image error or no English in the title have been excluded.

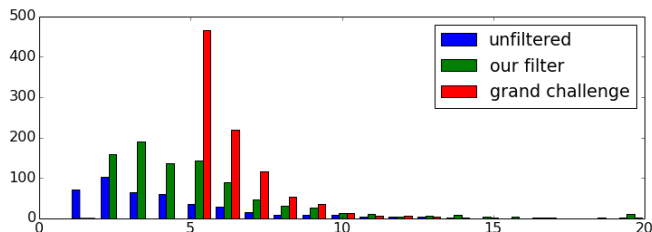


Figure 2: Caption length distributions for the different filtering methods. (Clipped at 20.)

For example, the same entity in an image could be referred to by using different synonyms. However, analyzing this automatically is extremely challenging, especially since the abundance of other kinds of variations make it challenging to even properly define style. Therefore, going into any details here is beyond the scope of this paper.

In general we assume that the purpose or intent the author might have for creating the caption has a strong effect on which particular route is taken in the process outlined above.

Also, context information (e.g. about the time or place of the image, names of people) is included in many of the captions. This is actually not surprising if we see image captioning as part of a multimodal communication act: In common social situations we can expect that people obey the *cooperative principle* which was introduced by Paul Grice in 1975 [4] and is well-known in the field of social science. The cooperative principle consists of 4 maxims, one of which says “Do not make your contribution more informative than is required” (second part of the *maxim of quantity*). Hence, if an image is shared and one can assume that the reader will see both the image and the caption at the same time (or at least the image not later than the title), then a mere description of the visible image contents would be redundant information and a violation of the cooperative principle.¹¹

5.2 Noise

As with any social media sampled datasets, YFCC100M contains a significant amount of noise. But the definition of noise depends on the task at hand. For this experiment we defined noise as the following classes and collected annotations for them. These classes included 1) images which were no longer available, 2) non-English

¹¹This is in general reasonable to assume for Flickr. However, there are cases where not both the image and the caption are visible. One obvious example would be if the viewer is a blind person. In such cases, merely “translating” the image into text does not violate Grice’s maxim of quantity.

titles and 3) gibberish titles which could not be understood by our annotators.

In Figure 1, images marked with a red boundary show some of the sample noise in the dataset. Table 1 lists out the annotated noisy images for each of the filters. We can see that the image errors are distributed more or less equally which is to be expected since filtering was only done at a caption level. Our filter was able to remove most of the non-English captions whereas in the grand challenge data annotators did not find any non-English caption at all. For the remaining captions, the percentages of captions that were not understood by the annotators are similar for the unfiltered data and both filters.

Our filter (using parameters we used for compiling the selection in this study) flags around 170 of the “clean captions” from the unfiltered YFCC100M as noise. We had a closer look at these captions and found that most of them are either single word captions (in total around 70 captions) or mostly consist of names, unusual words or non-letter parts. Some typical examples would be “2010-05-03 11-01-03 Rainbow Lorikeet - IMG_4857”, “Ech, malort.”, “cockortwo island” or “Castine, Maine”.

For all of the following analysis we exclude all items that have an image error, no English in the title or a title that was not understood by the majority of annotators. We refer to the remaining captions as “clean captions”. The number of clean captions for each data selection can be found in Table 1.

5.3 Comparing the Selections

To get an idea of the general structure of captions, we computed caption lengths and checked how many adjectives there are and how often adjectives are directly followed by a noun.

The caption length distributions differ significantly between the different selections (See Figure 2). This is important to keep in mind because this alone might cause shifts w.r.t. other aspects of the data.

We POS tagged all captions used for the study to analyze the usage of adjectives, using the python library NLTK¹² [10] for tokenization and POS tagging. All through, adjectives seem to be followed by nouns in 71 – 84% of the cases, where this fraction is highest for our filter and lowest for the grand challenge selection. Interestingly, the adjective-to-noun ratio is fairly low for both the unfiltered YFCC100M data and data filtered with our filter (both around 1 to 20), whereas in the grand challenge selection this ratio is close to 1 to 4.¹³

¹²<http://www.nltk.org/>

¹³We computed the same numbers again after removing all short captions but this did only marginally affect the numbers.

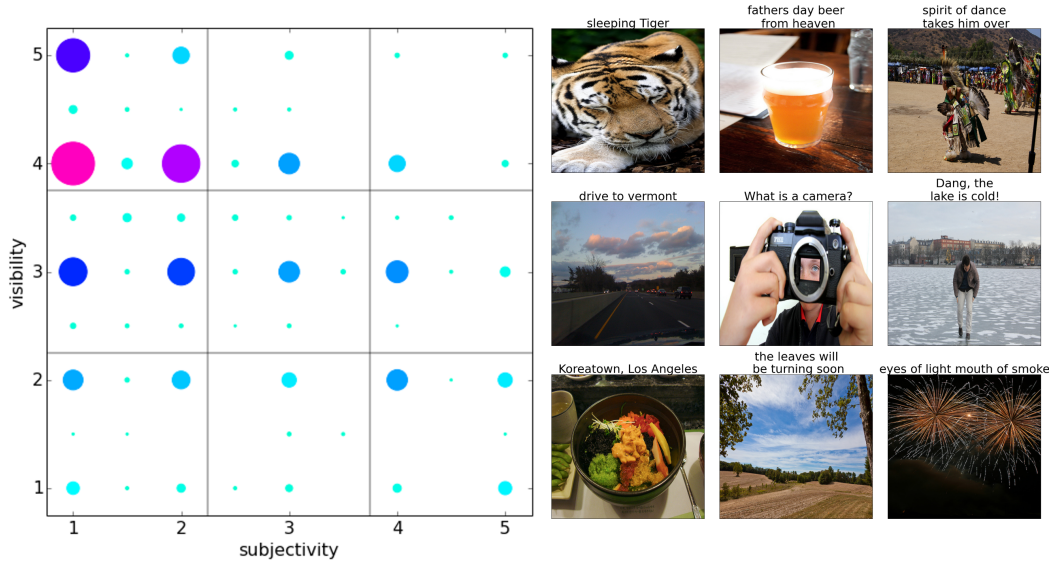


Figure 3: Illustration of the variation of visibility vs subjectivity. Each caption was assigned to a single visibility or subjectivity value based on the median values of the annotations for the respective field. The size of the circles represent the number of similar points on the graph. The majority of the captions lie in a high visibility, low subjectivity area. The figure on the right hand side shows the corresponding example for each of the 9 categories on the left.²

	factual	express	entertain	provoke	ambig
YFCC100M	85.61%	4.08%	5.76%	0%	4.56%
our filter	83.61%	5.50%	5.50%	0.34%	5.05%
GC	69.78%	8.39%	13.33%	0.43%	8.06%
aggregated	78.24%	6.43%	8.80%	0.31%	6.21%

Table 2: Intent distributions for the different selections, based on majority votes of the annotators on each of the selections. YFCC100M is the unfiltered selection and GC is the grand challenge selection.

For the intents we are reporting percentages based on majority votes. Note that one caption can belong to several intent categories in case there is a tie. For these cases we assign the caption to the category “ambiguous”. Table 2 lists out the distribution of intents across the three selections. It is interesting to note that factual descriptions dominate the intent category in all three selections. Still, this dominance is significantly weaker in the case of grand challenge captions which goes hand in hand with a higher number of expressive and entertaining captions. For our filter and the unfiltered YFCC100M data there are only small differences in intent distributions. As there are no captions with a majority vote for the intent “other”, we will ignore this intent category for the remainder of the paper.

The distributions of the responses for the other aspects are very comparable for the different selections, so for the rest of the analysis data of all selections is aggregated.

5.4 The Visibility-Subjectivity Space

Figure 3 shows the distribution of captions for different combinations of visibility and subjectivity (based on median values of all responses for each given image,caption pair).

To simplify interpretation, we separated this space into 9 boxes (as indicated by the lines in Figure 3). Statistics about each of these boxes can be found in Table 4. We can characterize these categories as follows:

High visibility, low subjectivity. This makes up the largest category. There are many captions that could be seen as titles in a classical sense. Very often, additional context information (such as entity names, naming rare objects/animals or place information) is included in the caption as well and some captions guide the viewer towards a certain interpretation of the image. Below we will analyze other aspects of this category.

High visibility, medium subjectivity. Typically some aspects of the image are interpreted in a subjective way in the caption. Often there is a clear relation to the image contents.

High visibility, high subjectivity. This combination is rather unusual. Usually the image content is described in a very subjective way and the captions can be of an artistic kind.

Medium visibility, low subjectivity. Captions here typically pick up on something visible in the image while giving additional background information.

Medium visibility, medium subjectivity. Quite heterogenous category. Some are rather subjective interpretations of the image, some merely give background information and some are rather artistic

Medium visibility, high subjectivity. Interesting category with lots of entertaining and expressive captions. Most of the captions can be seen as subjective interpretations of (parts of) the image.

Here, the relation to the image can be quite complex and captions should generally be treated as parts of a multimodal message.

Low visibility, low subjectivity. For the most part, these captions provide invisible background information (usually about the location where the image was taken, sometimes also about the time it was taken or about entities on the image - such as names of people).

Low visibility, medium subjectivity. Uncommon (smallest) and heterogenous category. Some captions give invisible background information, some relate to the image in rather complex ways and some “classic” titles.

Low visibility, high subjectivity. Intent here is for most cases either entertainment or expression. Majority of captions are subjective interpretations of the image content. Relation to the image is at times rather complex. There are also a few cases where there is no clear relation between the image and the title at all.

Since the high-visibility-low-subjectivity category is by far the largest and most existing work focuses on captions from this category, we had a closer look at it. In particular we wanted to see how captions in this category relate to captions from other existing datasets such as MS-COCO.

What we found is that even though a very large amount of captions are in this general category, most of the captions are at least slightly subjective or include information that is not easily recognizable in the image (even to most humans). For most image captioning datasets out there, all of the captions are visible and there is typically no subjectivity at all. So these types of captions would all reside in the top-left circle in Figure 3 which has around 10% of the captions.

When having a closer look at these captions in our dataset, however, we find that there are mainly the following subtypes:

- Classic titles (like “sleeping Tiger” or “Rails”)
- Background information that is somewhat visible, including naming uncommon things or animals, and named entities (such as “the great wall of China”)
- Highly descriptive titles (such as “Girls working in rice paddies”) do exist but are actually very rare.

It should not be very surprising that there are not many highly descriptive titles because this would in most circumstances be a violation of Grice’s maxim of quantity [4] (also see Section 5.1). The right part of Figure 3 displays one exemplary image-caption pair for each corresponding category on the left side.

5.5 Sentiment and Interactions

We did all sentiment calculation with NLTK vader and use the term sentiment to refer to the value returned as “compound sentiment”.

For finding out how sentiment and intent categories interact with appeal of the caption, we separated all captions into 5 groups based on the median values of preference responses and computed sentiment and intent distributions for all of these¹⁴. The results are shown in Table 3.

The results indicate that people generally prefer captions with a positive sentiment over neutral captions. Negativity in the sentiment seems to have an adverse effect on the appeal of the caption. For intents we can see that entertainment is more frequent in

¹⁴Captions with non-integer values have been assigned to both neighboring groups. This was done for 27 images in total.

median preference		1	2	3	4	5
sent.	average abs	0.08	0.10	0.10	0.10	0.14
	average	-0.08	0.03	0.04	0.05	0.11
	#negative	16.0%	8.9%	6.9%	7.1%	5.0%
	#positive	0%	14.4%	16.0%	18.0%	26.4%
intent	factual	76.0%	78.6%	81.2%	72.9%	67.7%
	provoke	8.0%	0.6%	0.2%	0.2%	0%
	express	4.0%	6.0%	5.8%	8.2%	8.1%
	entertain	8.0%	8.8%	5.9%	13.5%	18.2%
	ambiguous	4.0%	6.1%	6.9%	5.3%	6.1%
#captions		25	182	1351	608	99

Table 3: Interactions of annotator preference with estimated sentiment and intents of the captions. Numbers in a column with integer header “ n ” are calculated based on the set of captions with median preference value in the range $[n - 0.5, n + 0.5]$. The last row shows the total number of captions in the corresponding set. For sentiment, the average absolute sentiment (“average abs”), the average sentiment (“average”), the percentage of captions with a negative sentiment (“#negative”) and the percentage of captions with a positive sentiment (“#positive”) is shown. All sentiment values are estimated by using NLTK vader. The intent values are calculated from the crowd-sourcing data by using majority votes.

highly-rated captions. Most of the provoking captions apparently ended up receiving very low preference scores. For the other intent categories correlations are not as clear.

From manually looking at captions of very low and very high preference, we found that captions people do not like at all tend to have more negative sentiment (e.g. “fallen rear dump truck door”), be more generic (e.g. “Pictures from dig cam 10.06 026”), include “noisy” parts (e.g. “Pg 076i Historical Sites”) or lack information apart from names (e.g. “Jennie & Colette”). On the other side of the spectrum, people’s favorite captions typically relate to the image in a rather clear manner and often make the viewer see the image in some special way.

6 CONCLUSION

Different simple filters for noise removal have been described and their effects on several properties of the data were analyzed. We have seen theoretical explanations on potential sources of variability in image captions as well as quantitative ways of characterizing the image/caption space with the help of crowdsourcing. We found that even though lots of captions are highly visible and not very subjective, captions that merely describe obvious image contents are not very common in the wild. Our results also suggest that people tend to prefer caption with a positive sentiment over neutral captions which is an interesting finding, given the comparatively low number of captions with significant sentiment in our dataset.

In general we found that in the context of a photo-sharing platform, an image/caption pair is essentially a single multimodal message and it makes sense to model it as such if the goal is to enable

↑ visibility	↔ subjectivity	low subjectivity (median 1-2)	medium subjectivity (median 2.5-3.5)	high subjectivity (median 4-5)	any subjectivity
high visibility (median 4-5)	number of captions	1008 (45.04%)	115 (5.14%)	71 (3.17%)	1194 (53.35%)
	sentiment	0.06 / 0.02 / 5.4% / 10.9%	0.16 / 0.10 / 7.4% / 31.1%	0.30 / 0.23 / 7.3% / 53.6%	0.08 / 0.04 / 5.7% / 15.4%
	intent distribution	971	60	20	1042
	ambiguous factual entertainment provoke expression	20 1 6 10	20 0 13 22	1 11 17 22	60 2 36 54
medium visibility (median 2.5-3.5)	number of captions	359 (16.04%)	112 (5.00%)	128 (5.72%)	599 (26.76%)
	sentiment	0.06 / 0.02 / 5.1% / 10.6%	0.15 / 0.05 / 12.9% / 23.6%	0.22 / 0.10 / 15.3% / 35.0%	0.11 / 0.05 / 8.8% / 18.3%
	intent distribution	349	74	22	444
	ambiguous factual entertainment provoke expression	6 0 4	14 0 7 17	1 21 35 49	42 1 42 70
low visibility (median 1-2)	number of captions	203 (9.07%)	60 (2.68%)	182 (8.13%)	445 (19.88%)
	sentiment	0.07 / 0.03 / 4.5% / 12.0%	0.06 / 0.04 / 4.1% / 11.9%	0.19 / 0.05 / 15.6% / 25.6%	0.12 / 0.04 / 9.0% / 17.5%
	intent distribution	195	40	26	265
	ambiguous factual entertainment provoke expression	5 0 3 0	6 0 6 8	4 30 57 65	37 4 66 73
any visibility	number of captions	1570 (70.15%)	287 (12.82%)	381 (17.02%)	2238 (100.00%)
	sentiment	0.06 / 0.02 / 5.2% / 11.0%	0.13 / 0.07 / 8.8% / 24.0%	0.22 / 0.10 / 14.0% / 34.0%	0.10 / 0.04 / 7.2% / 16.6%
	intent distribution	1515	174	68	1751
	ambiguous factual entertainment provoke expression	31 1 9 14	40 0 26 47	6 62 109 136	139 7 144 197

Table 4: Numbers of captions in the visibility-subjectivity boxes. We assigned that captions to the different boxes based on the median values of the annotations for the respective question. The “numbers of captions” rows give the total numbers (or percentages respectively) of captions that fall within this category in the aggregated data set. The sentiment values are based on NLTK vader and include (in this order) the average absolute sentiment, the average sentiment, the percentage of captions with a negative sentiment and the percentage of captions with a positive sentiment. The intent distributions show how many frequent the different intents are within the given category (based on majority votes). The last row and last column contain aggregated statistics (combining all visibility levels or all subjectivity levels respectively).

machines to generate or understand image captions in a more natural way. At this point, we completely acknowledge that there is a long way to go until automatic image captioning can actually be considered solved. Lots of the caption types described in this study still seem out of reach and some simply do not seem to be addressed yet. In either case, we think that this paper moves some

of these aspects out of the periphery as it structures some of this unknown territory.

7 ACKNOWLEDGMENTS

This work was partially funded by the BMBF project Multimedia Opinion Mining (MOM: 01WI15002). The authors would like to



There is no image / only an error message.

1 Also mark this checkbox if a placeholder image is shown instead of the original image. (Indicated by a text such as 'This photo is no longer available.')

How much do you like the given image title? (required)

Not at all	1	2	3	4	5	Very much
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

There is no English in the title or there is no title at all.

1 Titles that only have non-English words and camera-generated titles such as img_123 should be rated as containing no English. If the whole title is written in a foreign language but contains names that are meaningful to English speakers (e.g. 'Reise mit Freunden durch London'), please also mark this box.

How subjective is the title? (required)

Purely objective	1	2	3	4	5	Purely subjective
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

1 Names, dates and other factual information are objective information, even if not visible in the image. Emotions, attitudes, opinions or other personal interpretations are considered to be subjective.

How much of the information given in the title can be seen in the image? (required)

Nothing	1	2	3	4	5	All of it
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

1 For example, suppose you are shown an image of a dog sitting on a crowded street. If the title is "dog", then it should be rated as 5 because all information in the title is visible. If the title is "on my way home with my dog", then it should have a rating of 2, because most of this information is invisible background information. Please note: Names (of people, animals, events or places) are considered to be visible only if they refer to something or someone very common and can be recognized in the given image.

Do you understand the title? (required)

- Yes
 No

In which situations would you most likely use such a title for the given image? (required)

- To entertain someone (e.g. in a humorous, witty, artistic or poetic way)
 To provoke someone (e.g. insult someone, tease someone or draw public attention)
 To report factual information
 To express emotions or an attitude
 Other

1 If you think that several situations are very plausible you can choose more than one option.

Figure 4: Screenshot of the crowdsourcing task as seen by an annotator.²

thank NVIDIA for support within the NVAIL program. The first

author received financial support from the Center for Cognitive Science (Kaiserslautern).

REFERENCES

- [1] Damian Borth, Christian Schulze, Adrian Ulges, and Thomas M Breuel. 2008. Navigator-similarity based browsing for image and video databases. In *Annual Conference on Artificial Intelligence*. Springer, 22–29.
- [2] Damian Borth, Adrian Ulges, and Thomas M Breuel. 2010. Relevance filtering meets active learning: improving web-based concept detectors. In *Proceedings of the international conference on Multimedia information retrieval*. ACM, 25–34.
- [3] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, and others. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1473–1482.
- [4] H. Paul Grice. 1975. Logic and Conversation. In *Speech acts*, Peter Cole (Ed.). Syntax and semantics, Vol. 3. Academic Press, New York, 41–58.
- [5] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2015. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571* (2015).
- [6] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. 2015. Real-time analysis and visualization of the YFCC100M dataset. In *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*. ACM, 25–30.
- [7] Tushar Karayil, Philipp Blandfort, Damian Borth, and Andreas Dengel. 2016. Generating Affective Captions using Concept And Syntax Transition Networks. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1111–1115.
- [8] Alireza Koochali, Sebastian Kalkowski, Andreas Dengel, Damian Borth, and Christian Schulze. 2016. Which Languages do People Speak on Flickr?: A Language and Geo-Location Study of the YFCC100m Dataset. In *Proceedings of the 2016 ACM Workshop on Multimedia COMMONS*. ACM, 35–42.
- [9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [10] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002). <http://arxiv.org/abs/cs.CL/0205028>
- [11] Karl Ni, Roger Pearce, Kofi Boakye, Brian Van Essen, Damian Borth, Barry Chen, and Eric Wang. 2015. Large-scale deep learning on the YFCC100M dataset. *arXiv preprint arXiv:1502.03409* (2015).
- [12] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.
- [13] Abebe Rorissa. 2010. A comparative study of Flickr tags and index terms in a general image collection. *Journal of the Association for Information Science and Technology* 61, 11 (2010), 2230–2242.
- [14] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. 2009. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 484–491.
- [15] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.
- [16] Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. 2011. Finding locations of flickr resources using language models and similarity search. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 48.
- [17] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555* (2014).
- [18] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. 2048–2057.
- [19] Sicheng Zhao, Hongxun Yao, Yue Gao, RongRong Ji, and Guiguang Ding. 2016. Continuous Probability Distribution Prediction of Image Emotions via Multi-Task Shared Sparse Regression. *PP* (10 2016), 1–1.
- [20] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. 2016. Predicting Personalized Emotion Perceptions of Social Images. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. ACM, New York, NY, USA, 1385–1394. DOI : <http://dx.doi.org/10.1145/2964284.2964289>

Introducing Concept And Syntax Transition Networks for Image Captioning

Philipp Blandfort
University of Kaiserslautern
Kaiserslautern, Germany.
philipp.blandfort@dfki.de

Tushar Karayil
University of Kaiserslautern
Kaiserslautern, Germany.
tushar.karayil@dfki.de

Damian Borth
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
damian.borth@dfki.de

Andreas Dengel
University of Kaiserslautern
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
andreas.dengel@dfki.de

ABSTRACT

The area of image captioning i.e. the automatic generation of short textual descriptions of images has experienced much progress recently. However, image captioning approaches often only focus on describing the content of the image without any emotional or sentimental dimension which is common in human captions. This paper presents an approach for image captioning designed specifically to incorporate emotions and feelings into the caption generation process. The presented approach consists of a Deep Convolutional Neural Network (CNN) for detecting Adjective Noun Pairs in the image and a novel graphical network architecture called “Concept And Syntax Transition (CAST)” network for generating sentences from these detected concepts.

Keywords

Auto Caption, Image Captioning

1. INTRODUCTION

With its exponential growth in the last two decades, the Internet has become a major source of information exchange across the world. Powered by new technologies and increased computational resources, web sites have become more visual and animated. Moreover, the world wide web is flooded with new images everyday, e.g. Instagram has reported an average of 80 million photo uploads a day.¹ Most of these images come with titles which can be generic (e.g. IMG_123), descriptive or give additional information that can not be directly seen in the image.

¹<https://www.instagram.com/press/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR'16, June 06 - 09, 2016, New York, NY, USA

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-4359-6/16/06...15.00

DOI: <http://dx.doi.org/10.1145/2911996.2930060>

Recently, there have been significant advances in generating descriptive image captions ([7],[8],[9]) but so far, the focus was on generating factual descriptive image captions like Microsoft COCO [3]. These datasets provide a rich textual description of images. However, these descriptions might not be representative for natural image captioning since human subjects were given clear instructions on writing the captions [3]. Such descriptions, although informative, constitute only a small subset of the different styles of captioning.

In contrast, large real-world datasets such as “Yahoo Flickr Creative Common 100 Million” (YFCC100M [12]) displays a huge variety of captioning styles but these titles can only be considered to be weak labels (cf. [13]): Here, captions can be descriptive, emotional or mention information that is not visible in the image.

Humans often tend to associate a sentiment with an image and express that in the caption. One such method is by using an emoji.² A richer use of text would be another way to express the associated sentiment with the caption.

It was shown by [1] that adjectives can add an emotional component to nouns and the resulting Adjective Noun Pairs can express the visual contents in the image. Hence we assume that incorporating adjectives into machine generated captions is one feasible way of adding an emotional component to the caption. To this end, we describe a model that is capable of generating subjective image captions and thereby going beyond factual image descriptions. We train and test our model on the YFCC100M database.

2. RELATED WORK

The available methods in linking images to text can be classified broadly into three categories.

The first set of methods is used to detect a triplet (eg. $\langle object, action, scene \rangle$) of scene elements in the image and convert them into sentences. Triplets provide a holistic idea of what is most important in the image and they are combined using various techniques to generate captions. [4], [8]

²<http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji>

use this approach and a template based system to generate the sentences after identifying the objects.

The second set of methods bring the images and sentences into a single multi dimensional space by converting each of them into vectors. Thereafter a set of distance measures are used to find the closest matching description of a given image ([5], [11]). [11] uses neural networks to map images and sentences into the same vector space. Although the above mentioned methods have shown promising results they cannot be used for generating novel descriptions. Hence the performance of these methods drop when there are new compositions of objects in a given image (even though individual objects have been observed during training).

The third set of methods which have shown the most promising results use a combination of Deep Convolutional Neural Networks (DCNN) for feature extraction and a Multimodal Recurrent Neural Networks (RNN) on top of it for text generation from the extracted features [14].

However, despite the promising results, approaches of the third kind suffer from several shortcomings:

- **Robustness:** RNNs heavily rely on suitable ground truth information. We argue that training an RNN on the YFCC100M would require expensive preprocessing since there are too many images with titles that can not be learned with current methods (e.g. because the relation to the visual content is not straight-forward) and could therefore disturb the training.
- **Transparency:** It is hardly possible to interpret what exactly is happening inside the RNN. The problem we see with this lack of transparency is that it forces you to treat the whole sentence generation part as single task that can not readily be broken down into distinct parts. This makes it hard to gain new insights about human language processing from the performance of the system or to incorporate new insights into the model. As a result, the performance depends heavily on the training data and modifications often have to be done by trial and error.

We apply *DeepSentiBank* as fixed visual concept extraction and generate captions from the features with the novel CAST network architecture instead of RNN in order to address the aforementioned shortcomings. This combination is quite robust and you can follow and influence all the steps from detected concepts to final sentence, giving you much more control and allowing for easier future changes of the architecture.

3. PRESENTED APPROACH

The presented system follows a pipeline approach consisting of the following steps:

1. **Visual concept extraction:** We process the image with *DeepSentiBank* to extract concepts including emotional cues.
2. **CAST network:** Generate ranked sentences from the detected concepts.
3. **Templates:** If the rankings from the network are below a threshold, we use a template-based approach to create sentences.

3.1 Data Preparation

The YFCC100M dataset contains user captioned Flickr images. Users' captions/tags often do not provide the appropriate data required to train classifiers and generate graphical language models. For example, the images often contain camera generated captions, generic titles, single word captions, locations as reported in [6]. Therefore it was important to extract the relevant Image-Caption pairs useful for model training.

We filter and remove all images with titles that are generic (e.g. IMG_1234.jpg), have less than 2 words or contain one or more non-English words.

After applying the filter, we end up with a training set consisting of 9.6 million images and a validation set consisting of 1.2 million images where the split into train and validation set is based on the user identifier present in the image meta-data. The cleaned up set still contains captions that are not directly related to the image, often providing extra information. But as this is natural in human image captioning, we deliberately keep this kind of data.

We build our graph (including word2vec model) on this data only, showing that our method works well with noisy real-world data without any sophisticated preprocessing.³

3.2 Visual Concept Extraction

To generate human like caption, the first step is to capture the emotional and visual contents from the image. The work published in [1], introduced Adjective Noun Pair (ANP) concepts able to describe images beyond visual content (e.g. "dog") by capturing positive or negative polarity (e.g. "cute dog" or "scary dog"). The resulting set of ANPs as trained by a deep convolution neural network is called *DeepSentiBank* and was published by [2]. This pairing of adjectives and nouns does also provide an insight into the general emotion associated with an analyzed image.

Processing the image with *DeepSentiBank* gives us a feature vector where each element corresponds to one ANP from a list of 2089 ANPs.

These 2089 ANPs contain 231 distinct adjectives and 424 nouns. This size is quite small when we want to generate different styles of sentences. To increase the size of vocabulary we generated a word2vec [10] model from all the training titles. (See next point.)

3.3 Concept And Syntax Transition (CAST) Network

The CAST network is a multi-directed graph where each node in the network represents a concept (i.e. noun, adjective, verb or adverb) connected to other concepts. It is generated in the following steps:

1. **Nodes:** For each content word with occurrence count greater than 40 in the training titles, we create a node. This leads to a vocabulary of over 21000 words. Additionally, we add a START and an END node.
2. **Similarity edges:** We train a word2vec model on all training sentences. word2vec maps words to a vector space of a given dimension (200 in our case) where words that are used in similar contexts are mapped to vectors that are close together in the target space.

³In principle it would be possible to train the visual concept detector on the YFCC100M data as well.

Under the assumption that semantically similar words are used similarly, this makes it possible to compute semantic distances between words.

For each node we add similarity edges to all nodes that have a word2vec similarity above some fixed threshold. This accounts for the possibility of replacing words by semantically similar words in the sentence generation process.

3. **Syntax edge:** For each sentence in the training titles we check how content words are connected. For each such connection that does not use another content word, a directed edge is created between the corresponding concept nodes. The connecting string (usually consisting of propositions, articles, etc.) is used as edge label and the total number of connection occurrences is annotated as edge weight.

These edges contain information about the syntax of the language and are used to connect different concepts.

The whole graph generation was done in less than 40h and the model can be used without any further training.

In CAST networks generating a sentence from a set of concepts is reduced to the problem of finding a path from the start to the end node through a set of activated nodes. Computing a list of such paths is done in a heuristic way and this list is then ranked by considering the weights of the included edges. The path with the highest score is then converted to a sentence in a straight-forward way, where similarity edges are used to substitute words. An illustration of a simple CAST network can be found in Figure 1.

In general, CAST networks provide a new possibility for the challenging task of generating sentences from an arbitrary sets of words. By using word similarity in the sentence generation process, they display a high degree of creativity, effectively extending the vocabulary. They do all this in a simple and transparent way which makes it easy to find the source of mistakes, allowing for systematic improvements or customization of the system in the future.

3.4 Template-based Approach

The idea of this approach is to use different templates of the kind “HUMAN with PROPERTY doing VERB on EVENT in LOCATION” to form sentences from a set of visual concepts that have been tagged by according category and are detected in the image.

For this we need:

- **Category tags:** We manually assigned category tags (e.g. “HUMAN” or “LOCATION”) to all nouns that occur in any ANPs.
- **Templates:** A few (5) templates based on these category tags were created manually. From that we automatically generated different template variations by removing parts of the template. (E.g. the variations of the above template would include “HUMAN doing VERB in LOCATION” and “HUMAN with PROPERTY on EVENT in LOCATION”.)

Sentences are now generated in the following steps:

1. **Input:** Given ANP scores from *DeepSentiBank*, we consider all ANPs that have a score above a fixed

threshold to create sentences from all suitable template variations. (If no score exceeds the threshold we take the ANP with highest confidence and return it as caption.)

2. **Ranking:** We rank all resulting sentences based on a scalar rating score that is computed for each sentence individually, using for the computation the *DeepSentiBank* scores of all ANPs that are present in the sentence.

3. **Output:** The sentence with the highest score is given as caption.

4. RESULTS

In order to evaluate the *humanness* factor of the generated caption, we selected 200 random images from our test set. These images were assigned two captions: The original caption present in the YFCC100M dataset and the caption generated by our method. Without informing the individual about the source of the two captions, we asked human subjects to choose one among the two captions which they thought were generated by a human. To compensate for the subjective bias in human evaluation, each image was shown to three different individuals and the opinion of the majority was decided as the final result for that image.

We report that 31.5% of the captions generated by our method were reported as more human-like in comparison to the original caption by at least two subjects. In 62.5% of images at least one subject chose our caption over the original one. These results are encouraging. The generated captions often read naturally and convey emotions. The creativity and subjectivity that is displayed in some of the captions is very entertaining. Figure 2 shows a small selection of titles generated by our method.

5. CONCLUSIONS

We presented an approach of combining the top Adjective Noun Pairs detected in an image with a graphical model to form captions that are not only descriptive but also carry subjective meaning. A human evaluation of our method on a subset of the YFCC100M dataset we often obtain natural image captions.

To improve the existing model and to get the caption quality closer to human levels we are planning to extend our work by incorporating the following points: The grammar of the whole sentence needs to be given more weight. We are currently working on an additional ranking mechanism to take that into account.

Also, so far the confidences of the detector are only respected in the thresholding and then discarded. We plan to either modify the network traversing algorithm such that it also respects the concept scores or respect the scores in the final ranking of the proposed sentences. We also want to use additional concept detectors to get more different sentences from the network and optimize the whole network on more data.

6. ACKNOWLEDGMENTS

This work was partially funded by the BMBF project Multimedia Opinion Mining (MOM: 01WI15002).

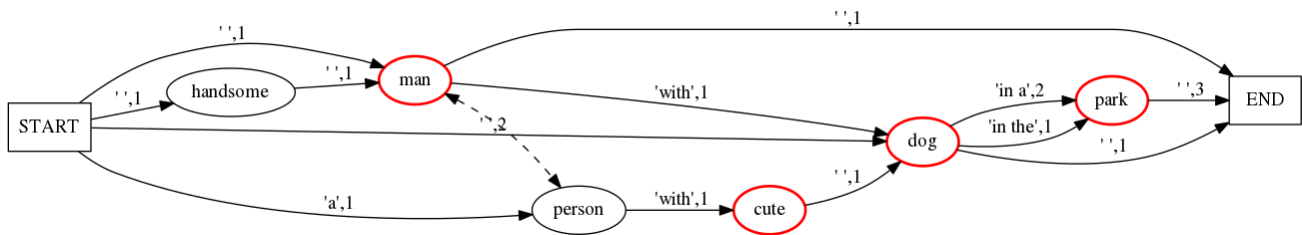
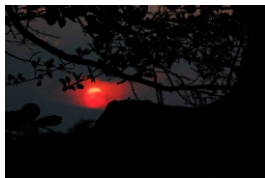


Figure 1: Example of a CAST network generated from the titles “handsome man”, “a person with cute dog”, “dog in a park”, “dog in the park” and “man with dog in a park”. The dashed line indicates a similarity edge (and is in this toy example not generated from the given sentences). If the red nodes denote the activated concepts, the resulting sentence would be “person with cute dog in a park”. (Substituting “man” by “person” because a similarity edge was traversed.)



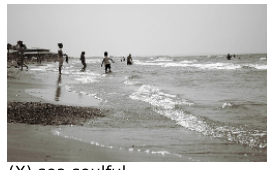
() fire in the sky, fire island
(X) nightfall and trees



() fruit op
(X) mucky and tired baby



() cloud claws
(X) violent storm clouds



(X) sea soulful
() cruel sea waves on the beach



(X) burning man
() amazing sky highway



(X) games convention storm trooper
() violent crime with an audience

Figure 2: Qualitative results of our approach for images of the YFCC100M dataset. The captions in black are the ground truth titles, in blue we have captions produced by the combination of *DeepSentiBank* and CAST. The “X” marks indicate which caption the majority of people in our evaluation experiment believed to be created by a human.

7. REFERENCES

- [1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *ACM Int. Conf. on Multimedia (ACM MM)*, 2013.
- [2] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.
- [3] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.
- [4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision—ECCV 2010*, pages 15–29. Springer, 2010.
- [5] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.
- [6] S. Kalkowski, C. Schulze, A. Dengel, and D. Borth. Real-time analysis and visualization of the yfcc100m dataset. In *MM COMMOMS Workshop*, 2015.
- [7] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.
- [8] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. ACL, 2011.
- [9] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
- [10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.
- [11] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the ACL*, 2:207–218, 2014.
- [12] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [13] A. Ulges, D. Borth, and T. M. Breuel. Visual concept learning from weakly labeled web videos. In *Video Search and Mining*, pages 203–232. Springer, 2010.
- [14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

Generating Affective Captions using Concept And Syntax Transition Networks

Tushar Karayil*
University of Kaiserslautern
Kaiserslautern, Germany.
tushar.karayil@dfki.de

Philipp Blandfort*
University of Kaiserslautern
Kaiserslautern, Germany.
philipp.blandfort@dfki.de

Damian Borth
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
damian.borth@dfki.de

Andreas Dengel
University of Kaiserslautern
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
andreas.dengel@dfki.de

ABSTRACT

The area of image captioning i.e. the automatic generation of short textual descriptions of images has experienced much progress recently. However, image captioning approaches often only focus on describing the content of the image without any emotional or sentimental dimension which is common in human captions. This paper presents an approach for image captioning designed specifically to incorporate emotions and feelings into the caption generation process. The presented approach consists of a Deep Convolutional Neural Network (CNN) for detecting Adjective Noun Pairs in the image and a graphical network architecture called “Concept And Syntax Transition (CAST)” network for generating sentences from these detected concepts.

Keywords

Image Captioning, Sentence Generation, Auto Caption, Sentiment, Emotion

1. INTRODUCTION

With its exponential growth in the last two decades, the Internet has become a major source of information exchange across the world. Powered by new technologies and increased computational resources, web sites have become more visual and animated. Moreover, the world wide web is flooded with new images everyday, e.g. Instagram has reported an average of 80 million¹ photo uploads a day. Most of these images come with titles which can be generic (e.g. IMG_123), descriptive or give additional information that can not be

*Equal contribution from both authors

¹<https://www.instagram.com/press/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '16, October 15-19, 2016, Amsterdam, Netherlands

© 2016 ACM. ISBN 978-1-4503-3603-1/16/10...\$15.00

DOI: <http://dx.doi.org/10.1145/2964284.2984070>

directly seen in the image but can be used in an multimodal retrieval setup [17].

Recently, there have been significant advances in generating descriptive image captions ([9],[10],[11]) with focus on generating *factual* descriptive image captions. Available datasets like Microsoft-COCO [4] provide a rich textual description of images. In contrast, large real-world datasets such as “Yahoo Flickr Creative Commom 100 Million” (YFCC100M [15]) display a huge variety of captioning styles but these titles can only be considered to be weak labels (cf. [16]): Here, captions can be descriptive, emotional or mention information that is not visible in the image.

Humans often tend to associate a sentiment with an image and express that in the caption [19]. One such method is by using a emoji². In addition, a richer use of text would be another way to express the associated sentiment with the caption.

It was shown by [2] that adjectives can add an emotional component to nouns and the resulting Adjective Noun Pairs (ANP) can express the visual contents in the image. Hence we assume that incorporating adjectives into machine generated captions is one feasible way of adding an emotional component to the caption.

To this end, we describe a model that is capable of generating subjective image captions and thereby going beyond factual image descriptions. We train and test our model on the YFCC100M dataset in the context of the caption prediction task of the *ACM Multimedia 2016 Grand Challenge*.

2. RELATED WORK

The available methods in linking images to text can be classified broadly into three categories.

The first set of methods is used to detect a triplet (eg. $\langle object, action, scene \rangle$) of scene elements in the image and convert them into sentences. Triplets provide a holistic idea of what is most important in the image and they are combined using various techniques to generate captions. [5], [10] use this approach and a template based system to generate

²<http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji>

the sentences after identifying the objects. Although, template based approaches generate a correct description, they cannot create a novel or unique caption as they are bound by the structure of the template used.

The second set of methods bring the images and sentences into a single multi dimensional space by converting each of them into vectors. Thereafter a set of distance measures are used to find the closest matching description of a given image ([6], [14]). [14] uses neural networks to map images and sentences into the same vector space. The main disadvantage here is that images which resemble each other closely in color space need not be semantically related. Therefore the generated captions can be completely unrelated to the given image.

The third set of methods which have shown the most promising results use a combination of Deep Convolutional Neural Networks (DCNN) for feature extraction and a Recurrent Neural Networks (RNN) on top of it for text generation from the extracted features ([18], [12]).

To the best of our knowledge, *SentiCap* [12] is the only neural network based approach for image captioning that deliberately incorporates sentiment into the produced captions. However, this approach relies on ground truth captions with known sentiment values which are not part of the Grand Challenge training data. Moreover, we doubt that using the captions they used would give good results in our case since the style of these captions is quite different from typical YFCC100M captions.

The strong dependency on the training data of RNN based models might in general be problematic when working with YFCC100M data because of the variety of caption types. Since such models are typically trained in an end-to-end fashion resulting issues can be extremely hard to fix.

In order to avoid these problems, we use an existing visual concept detector and generate captions from the extracted concepts with the CAST network architecture instead of RNN. This combination is quite robust and you can follow and influence all the steps from detected concepts to final sentence, giving you much more control and allowing for easier future changes of the architecture.

3. PRESENTED APPROACH

The presented system follows a pipeline approach consisting of the following steps:

1. **Visual concept extraction:** Extract subjective visual concepts from the given image.
2. **CAST network:** Generate ranked sentences from the detected concepts.
3. **Templates:** In case the network can not build any sentence from the given concepts we use a template-based back-up approach to create sentences.

3.1 Data Preparation

The data for the ACM Multimedia Grand Challenge 2016 was provided in the following format:

- Training: 1287522 train images from the YFCC100M dataset with captions that contain at least one word from English Dictionary. It is known that such images are weakly labeled[16]. In addition to the image identifier, the user identifier is also provided.
- Testing: 36884 test images with image-id and user-id

3.2 Visual Concept Extraction

To generate human-like captions, the first step is to capture emotional and visual contents from the image. The work published in [2] introduced Adjective Noun Pair (ANP) concepts able to describe images beyond visual content (e.g. “dog”) by capturing positive or negative polarity (e.g. “cute dog” or “scary dog”). This pairing of adjectives and nouns does also provide an insight into the general emotion associated with an analyzed image.

The DCNN *DeepSentiBank* [3] was trained to detect 2089 such ANPs (with 231 distinct adjectives and 424 nouns).

3.3 Concept And Syntax Transition (CAST) Network

The CAST network is a multi-directed graph where each node in the network represents a concept (i.e. noun, adjective, verb or adverb) connected to other concepts. It is generated in the following steps:

1. **Nodes:** For each content word with occurrence count greater than 40 in the training titles, we create a node. This leads to a vocabulary of over 21000 words. Additionally, we add a START and an END node.
2. **Similarity edges:** The vocabulary of *DeepSentiBank* is quite small if we want to generate different styles of sentences. In order to effectively increase the vocabulary size we train a word2vec [13] model on all training sentences. word2vec maps words to a vector space of a given dimension (200 in our case) where words that are used in similar contexts are mapped to vectors that are close together in the target space. Under the assumption that semantically similar words are used similarly, this makes it possible to compute semantic distances between words. For each node we add similarity edges to all nodes that have a word2vec similarity above some fixed threshold. This accounts for the possibility of replacing words by semantically similar words in the sentence generation process.
3. **Syntax edges:** For each sentence in the training titles we check how content words are connected. For each such connection that does not use another content word, a directed edge is created between the corresponding concept nodes. The connecting string (usually consisting of propositions, articles, etc.) is used as edge label and the total number of connection occurrences is annotated as edge weight. These edges contain information about the syntax of the language and are used to connect different concepts.

The whole graph generation was done in less than 40h and the model can be used without any further training. Generating a sentence from a set of concepts is reduced to the problem of finding a path from the start to the end node through a set of activated nodes. We compute a list of such paths in a heuristic way. For each path a score is then computed by summing up the normalized weights of the included edges. Finally, the paths are ranked by these scores and the path with highest score is converted to a sentence in a straight-forward way, where similarity edges

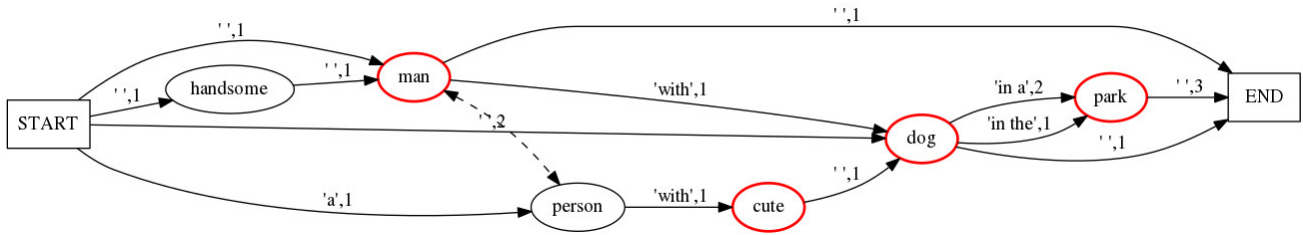


Figure 1: Example of a CAST network generated from the titles “handsome man”, “a person with cute dog”, “dog in a park”, “dog in the park” and “man with dog in a park”. The dashed line indicates a similarity edge (and is in this toy example not generated from the given sentences). If the red nodes denote the activated concepts, the resulting sentence would be “person with cute dog in a park”. (Substituting “man” by “person” because a similarity edge was traversed.)

are used to substitute words.³ An illustration of a simple CAST network can be found in Figure 1.

In general, CAST networks provide a new possibility for the challenging task of generating sentences from an arbitrary set of words. By using word similarity in the sentence generation process, they display a high degree of creativity, effectively extending the vocabulary. They do all this in a simple and transparent way which makes it easy to find the source of mistakes, allowing for systematic improvements or customization of the system in the future.

3.4 Template-based Approach

The idea of this approach is to use different templates of the kind “HUMAN with PROPERTY doing VERB on EVENT in LOCATION” to form sentences from a set of visual concepts that have been tagged by according category and are detected in the image.

For this we need:

- **Category tags:** We manually assigned category tags (e.g. “HUMAN” or “LOCATION”) to all nouns that occur in any ANPs.
- **Templates:** A few (5) templates based on these category tags were created manually. From that we automatically generated different template variations by removing parts of the template. (E.g. the variations of the above template would include “HUMAN doing VERB in LOCATION” and “HUMAN with PROPERTY on EVENT in LOCATION”).

Sentences are now generated in the following steps:

1. **Input:** Given ANP scores from *DeepSentiBank*, we consider all ANPs that have a score above a fixed threshold to create sentences from all suitable template variations where every category tag is substituted by an ANP of that category. (If no score exceeds the threshold we take the ANP with highest confidence and return it as caption.)
2. **Ranking:** For each resulting sentence we first compute the mean m and the sum s over the *DeepSentiBank* scores of all ANPs that are present in the sen-

tence. The weighted sum $0.3 \cdot m + 0.5 \cdot s$ is then used as final score for the ranking.⁴

3. **Output:** The sentence with the highest score is given as caption.

4. RESULTS AND ANALYSIS

CAST achieved a BLEU score of **0.04** at the leaderboard evaluation. (For the template-based approach the score was 0.02.)

Automatic evaluation of generated captions against the ground truth is a challenging task. One of the main reasons for this difficulty is that two captions which are syntactically different can mean the same thing. For example, an image of a rose could be captioned with “a rose” or “cute and pretty flower”. Although both of these captions suit the image, from a machine’s point of view they are completely different. Moreover, the scores are affected by the length of the generated captions. A low score might partly be due to the difference in caption length distributions of the generated captions in comparison to the ground truth data. By comparing the corresponding caption length histograms we found that in our case there is indeed a clear discrepancy. These histograms are shown in Figure 2.

The aforementioned points show that metrics like BLEU might fail to capture the correctness of the generated captions. With the absence of metrics addressing these issues, researchers [12, 8] have resorted to human evaluation of the generated captions.

To prove this point further, we calculated the FC7 vectors (last fully connected layer of *DeepSentiBank* [3]) for 300000 images from the training set and the whole test set. For each FC7 vector of a test image we found the nearest FC7 vector in the train set (using Euclidean distance) and output the title of the corresponding training image as the caption for the test image. As expected, this baseline method gave a higher score of **0.06**. Even though this method gave us a better score, we found that the generated sentences were less related to the images than captions created with CAST networks.

The generated captions from CAST often read naturally and conveyed emotions. Figure 3 shows a small selection of titles generated by our method in comparison to captions created with the Nearest Neighbour baseline approach. In order to get further insights into the emotional aspects of the

³Due to the graph size a full search is not feasible and hence it might happen that no path is found in which case the template-based approach serves as back-up method.

⁴Suitable multipliers were determined empirically.

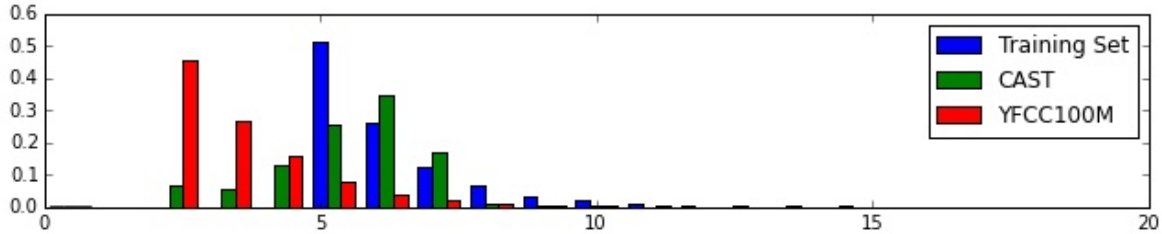


Figure 2: Normalized histograms of the caption lengths over three sets of captions: The ACMMM Grand Challenge 2016 training data (“Training Set”), CAST and a selection of 9 Million YFCC100M english titles obtained by simple filtering as described in [1]. (The filtered YFCC100M titles were included to get a rough estimate of the general distribution of user provided captions.)

Source	#Captions	Absolute Sentiment
CAST	36884	0.3
Training Set	1200000	0.14
MS-COCO	203450	0.07

Table 1: Average absolute sentiment value of captions from different sources. The ACMMM Grand Challenge 2016 training data (“Training Set”) which is taken from the YFCC100M displays a higher degree of sentiment than MS-COCO captions. CAST captions have an even higher absolute sentiment value which is twice that of the training set.

captions, we used VADER sentiment analysis tools [7] to calculate the average absolute sentiment value of the captions generated by CAST, the Grand Challenge training data and the MS-COCO dataset. It can be seen in Table 1 that by the usage of ANPs together with CAST a significant amount of sentiment could be added to the captions.

5. CONCLUSIONS

We presented an approach of combining the top Adjective Noun Pairs detected in an image with a graphical model to form captions that are not only descriptive but also carry subjective meaning. We also pointed out limitations of evaluation metrics like BLEU for captions by showing how baseline methods like Nearest Neighbours with FC7 can yield higher BLEU score even without generating any novel sentences.

To improve the existing model further and to get the caption quality closer to human levels we are planning to extend our work by incorporating additional ranking mechanisms and concept detectors.

6. ACKNOWLEDGMENTS

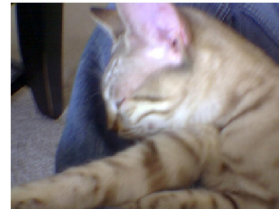
This work was partially funded by the BMBF project Multimedia Opinion Mining (MOM: 01WI15002). The second author would like to thank the Center for Cognitive Science (Kaiserslautern) for financial support.



(N) and so the drinking starts
(C) kind of looks like a hardcore



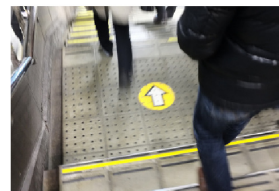
(N) it runs in the family
(C) portrait with hair



(N) jumping on the turtles back
(C) cat that is sleepy and lazy



(N) jogging through the seasons spring
(C) grass by the river in winter ice



(N) come with me if you want to live
(C) looks like a fluffy toy



(N) mountains of glacier national park
(C) misty mountains in valley road

Figure 3: Example results of our approach for images of the test dataset. The captions from CAST network are marked with “C” and the captions retrieved by FC7 nearest neighbours are marked with “N”.

7. REFERENCES

- [1] P. Blandfort, T. Karayil, D. Borth, and A. Dengel. Introducing concept and syntax transition networks for image captioning. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 385–388. ACM, 2016.
- [2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *ACM Int. Conf. on Multimedia (ACM MM)*, 2013.
- [3] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. DeepSentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.
- [4] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.
- [5] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision—ECCV 2010*, pages 15–29. Springer, 2010.
- [6] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.
- [7] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text, 2014.
- [8] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *arXiv preprint arXiv:1511.07571*, 2015.
- [9] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.
- [10] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics, 2011.
- [11] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
- [12] A. Mathews, L. Xie, and X. He. SentiCap: generating image descriptions with sentiments. In *The Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, Phoenix, USA, feb 2016.
- [13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.
- [14] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.
- [15] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [16] A. Ulges, D. Borth, and T. M. Breuel. Visual concept learning from weakly labeled web videos. In *Video Search and Mining*, pages 203–232. Springer, 2010.
- [17] A. Ulges, C. Schulze, D. Borth, and A. Stahl. Pornography detection in video benefits (a lot) from a multi-modal approach. In *Proceedings of the 2012 ACM international workshop on Audio and multimedia methods for large-scale video analysis*, pages 21–26. ACM, 2012.
- [18] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.
- [19] R. B. Zajonc. Feeling and thinking: Preferences need no inferences. *American psychologist*, 35(2):151, 1980.

Multimodal Social Media Analysis for Gang Violence Prevention

Philipp Blandfort,^{1,2} Desmond U. Patton,³ William R. Frey,³ Svebor Karaman,³
 Surabhi Bhargava,³ Fei-Tzin Lee,³ Siddharth Varia,³ Chris Kedzie,³
 Michael B. Gaskell,³ Rossano Schifanella,⁴ Kathleen McKeown,³ Shih-Fu Chang³

¹DFKI, ²TU Kaiserslautern, ³Columbia University, ⁴University of Turin
 philipp.blandfort@dfki.de, schifane@di.unito.it, {kedzie, kathy}@cs.columbia.edu
 {dp2787, w.frey, svebor.karaman, sb4019, fl2301, sv2504, mbg2174, sc250}@columbia.edu

Abstract

Gang violence is a severe issue in major cities across the U.S. and recent studies have found evidence of social media communications that can be linked to such violence in communities with high rates of exposure to gang activity. In this paper we partnered computer scientists with social work researchers, who have domain expertise in gang violence, to analyze how public tweets with images posted by youth who mention gang associations on Twitter can be leveraged to automatically detect psychosocial factors and conditions that could potentially assist social workers and violence outreach workers in prevention and early intervention programs. To this end, we developed a rigorous methodology for collecting and annotating tweets. We gathered 1,851 tweets and accompanying annotations related to visual concepts and the *psychosocial codes: aggression, loss, and substance use*. These codes are relevant to social work interventions, as they represent possible pathways to violence on social media. We compare various methods for classifying tweets into these three classes, using only the text of the tweet, only the image of the tweet, or both modalities as input to the classifier. In particular, we analyze the usefulness of mid-level visual concepts and the role of different modalities for this tweet classification task. Our experiments show that individually, text information dominates classification performance of the *loss* class, while image information dominates the *aggression* and *substance use* classes. Our multimodal approach provides a very promising improvement (18% relative in mean average precision) over the best single modality approach. Finally, we also illustrate the complexity of understanding social media data and elaborate on open challenges. The annotated dataset will be made available for research with strong ethical protection mechanism.

1 Introduction

Gun violence is a critical issue for many major cities. In 2016, Chicago saw a 58% surge in gun homicides and over 4,000 shooting victims, more than any other city comparable in size (Kapustin et al. 2017). Recent data suggest that gun violence victims and perpetrators tend to have gang associations (Kapustin et al. 2017). Notably, there were fewer homicides originating from physical altercations in 2016 than in the previous year, but we have little empirical evidence explaining why. Burgeoning social science research indicates

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

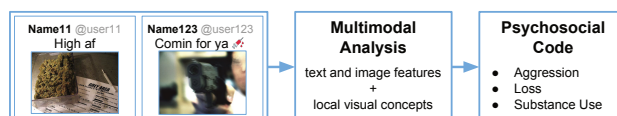


Figure 1: We propose a multimodal system for detecting psychosocial codes of social media tweets related to gang violence.

that gang violence may be exacerbated by escalation on social media and the “digital street” (Lane 2016) where exposure to aggressive and threatening text and images can lead to physical retaliation, a behavior known as “Internet banging” or “cyberbanging” (Patton, Eschmann, and Butler 2013).

Violence outreach workers present in these communities are thus attempting (Cit 2017a) to prioritize their outreach around contextual features in social media posts indicative of offline violence, and to try to intervene and de-escalate the situation when such features are observed. However, as most tweets do not explicitly contain features correlated with pathways of violence, an automatic or semi-automatic method that could flag a tweet as potentially relevant would lower the burden of this task. The automatic interpretation of tweets or other social media posts could therefore be very helpful in intervention, but quite challenging to implement for a number of reasons, e.g. the informal language, the African American Vernacular English, and the potential importance of context to the meaning of the post. In specific communities (e.g. communities with high rates of violence) it can be hard even for human outsiders to understand what is actually going on.

To address this challenge, we have undertaken a first multimodal step towards developing such a system that we illustrate in Figure 1.¹ Our major contributions lie in innovative application of multimedia analysis of social media in practical social work study, specifically covering the following components:

¹Note that the “tweets” in Figure 1 were created for illustrative purpose using Creative Commons images from Flickr and are NOT actual tweets from our corpus. Attributions of images in Figure 1, from left to right: “IMG.0032.JPG” by sashimikid, used under CC BY-NC-ND 2.0, “gun” by andrew_xjy, used under CC BY-NC-ND 2.0.

- We have developed a rigorous framework to collect context-correlated tweets of gang-associated youth from Chicago containing images, and high-quality annotations for these tweets.
- We have teamed up computer scientists and social work researchers to define a set of visual concepts of interest.
- We have analyzed how the psychosocial codes *loss*, *aggression*, and *substance use* are expressed in tweets with images and developed methods to automatically detect these codes, demonstrating a significant performance gain of 18% by multimodal fusion.
- We have trained and evaluated detectors for the concepts and psychosocial codes, and analyzed the usefulness of the local visual concepts, as well as the relevance of image vs. text for the prediction of each code.

2 Related Work

The City of Chicago is presently engaged in an attempt to use an algorithm to predict who is most likely to be involved in a shooting as either a victim or perpetrator (Cit 2017b); however, this strategy has been widely criticized due to lack of transparency regarding the algorithm (Schmidt 2018; Sheley 2017) and the potential inclusion of variables that may be influenced by racial biases present in the criminal justice system (e.g. prior convictions) (BBC 2017; Nellis et al. 2008).

In (Gerber 2014), Gerber uses statistical topic modeling on tweets that have geolocation to predict how likely 20 different types of crimes are to happen in individual cells of a grid that covers the city of Chicago. This work is a large scale approach for predicting future crime locations, while we detect codes in individual tweets related to future violence. Another important difference is that (Gerber 2014) is meant to assist criminal justice decision makers, whereas our efforts are community based and have solid grounding in social work research.

Within text classification, researchers have attempted to extract social events from web data including detecting police killings (Keith et al. 2017), incidents of gun violence (Pavlick et al. 2016), and protests (Hanna 2017). However, these works primarily focus on extracting events from news articles and not on social media and have focused exclusively on the text, ignoring associated images.

The detection of local concepts in images has made tremendous progress in recent years, with recent detection methods (Girshick 2015; Ren et al. 2017; Dai et al. 2016; Liu et al. 2016; Redmon et al. 2016) leveraging deep learning and efficient architecture enabling high quality and fast detections. These detection models are usually trained and evaluated on datasets such as the PascalVOC (Everingham et al. 2010) dataset and more recently the MSCOCO (Lin et al. 2014) dataset. However, the classes defined in these datasets are for generic consumer applications and do not include the visual concepts specifically related to gang violence, defined in section 3.2. We therefore need to define a lexicon of gang-violence related concepts and train own detectors for our local concepts.

The most relevant prior work is that of (Blevins et al. 2016). They predict *aggression* and *loss* in the tweets of Gakirah Barnes and her top communicators using an extensive set of linguistic features, including mappings of African American vernacular English and emojis to entries in the Dictionary of Affective Language (DAL). The linguistic features are used in a linear Support Vector Machine (SVM) to make a 3-way classification between *loss*, *aggression*, and *other*. In this paper we additionally predict the presence of substance use, and model this problem as three binary classification problems since multiple codes may simultaneously apply. We also explore character and word level Convolutional Neural Network (CNN) classifiers, in addition to exploiting image features and their multimodal combinations.

3 Dataset

In this section we detail how we have gathered and annotated the data used in this work.

3.1 Obtaining Tweets

Working with community social workers, we identified a list of 200 unique users residing in Chicago neighborhoods with high rates of violence. These users all suggest on Twitter that they have a connection, affiliation, or engagement with a local Chicago gang or crew. All of our users were chosen based on their connections to a seed user, Gakirah Barnes, and her top 14 communicators in her Twitter network (top communicators were statistically calculated by most mentions and replies to Gakirah Barnes). Gakirah was a self-identified gang member in Chicago, before her death in April, 2014. Additional users were collected using snowball sampling techniques (Atkinson and Flint 2001). Using the public Twitter API, in February 2017 we scraped all obtainable tweets from this list of 200 users. For each user we then removed all retweets, quote tweets and tweets without any image, limiting the number of remaining tweets per user to 20 to avoid most active users being overrepresented. In total the resulting dataset consists of 1,851 tweets from 173 users.

3.2 Local Visual Concepts

To extract relevant information in tweet images related to gang violence, we develop a specific lexicon consisting of important and unique visual concepts often present in tweet images in this domain. This concept list was defined through an iterative process involving discussions between computer scientists and social work researchers. We first manually went through numerous tweets with images and discussed our observations to find which kind of information could be valuable to detect, either for direct detection of “interesting” situations but also for extracting background information such as affiliation to a specific gang that can be visible from a tattoo. Based on these observations we formulated a preliminary list of visual concepts. We then collectively estimated utility (how useful is the extraction of the concept for gang violence prevention?), detectability (is the concept visible and discriminative enough for automatic detection?), and observability for reliable annotation (can we expect to

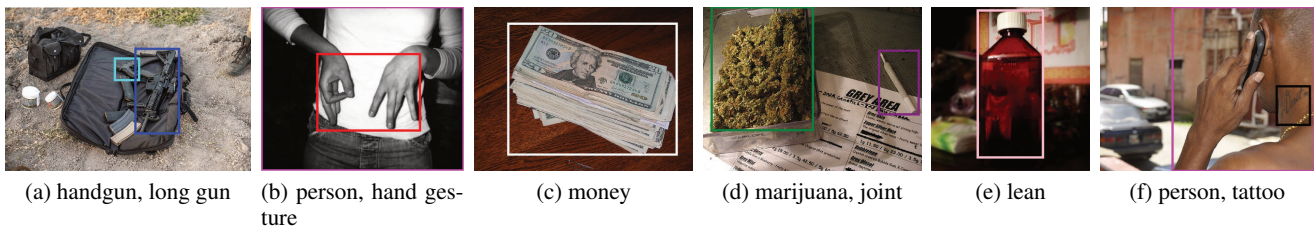


Figure 2: Examples of our gang-violence related visual concepts annotated on Creative Commons images downloaded from Flickr.

obtain a sufficient number of annotations for the concept?), in order to refine this list of potential concepts and obtain the final lexicon.

Our interdisciplinary collaboration helped to minimize the risk of overlooking potentially important information or misinterpreting behaviors that are specific to this particular community. For example, on the images we frequently find people holding handguns with an extended clip and in many of these cases the guns are held at the clip only. The computer scientists of our team did not pay much attention to the extended clips and were slightly confused by this way of holding the guns, but then came to learn that in this community an extended clip counts as a sort of status symbol, hence this way of holding is meant to showcase a common status symbol. Such cross-disciplinary discussions lead to inclusion of concepts such as *tattoos* and separation of concepts to *handgun* and *long gun* in our concept lexicon.

From these discussions we have derived the following set of local concepts (in image) of interest:

- General: *person, money*
- Firearms: *handgun, long gun*
- Drugs: *lean, joint, marijuana*
- Gang affiliation: *hand gesture, tattoo*

This list was designed in such a way that after the training process described above, it could be further expanded (e.g., by specific hand gestures or actions with guns). We give examples of our local concepts in Figure 2.²

3.3 Psychosocial Codes

Prior studies (Blevins et al. 2016; Patton et al. 2017) have identified *aggression*, *loss* and *substance use* as emergent themes in initial qualitative analysis that were associated with Internet banging, an emerging phenomenon of gang affiliates using social media to trade insults or make violence threats. Aggression was defined as posts of communication

²Attributions of Figure 2, from left to right: “GUNS” by djlinlovely, used under CC BY-NC-ND 2.0, “my sistah the art gangstah” by barbietron, used under CC BY-NC 2.0, “Money” by jollyuk, used under CC BY 2.0, “IMG.0032.JPG” by sashimikid, used under CC BY-NC-ND 2.0, “#codeine time” by amayzun, used under CC BY-NC-ND 2.0, “G Unit neck tattoo, gangs Trinidad” by bbcworldservice, used under CC BY-NC 2.0. Each image has been modified to show the bounding boxes of the local concepts of interest present in it.

that included an insult, threat, mentions of physical violence, or plans for retaliation. Loss was defined as a response to grief, trauma or a mention of sadness, death, or incarceration of a friend or loved one. Substance use consists of mentions, and replies to images that discuss or show any substance (e.g., marijuana or a liquid substance colloquially referred to as “lean”, see example in Figure 2) with the exception of cigarettes and alcohol.

The main goal of this work is to automatically detect a tweet that can be associated with any or multiple of these three psychosocial codes (*aggression*, *loss* and *substance use*) exploiting both textual and visual content.

3.4 Annotation

The commonly used annotation process based on crowd sourcing like Amazon Mechanical Turk is not suitable due to the special domain-specific context involved and the potentially serious privacy issues associated with the users and tweets.

Therefore, we adapted and modified the Digital Urban Violence Analysis Approach (DUVAA) (Patton et al. 2016; Blevins et al. 2016) for our project. DUVAA is a contextually-driven multi-step qualitative analysis and manual labeling process used for determining meaning in both text and images by interpreting both on- and offline contextual features. We adapted this process in two main ways. First, we include a step to uncover annotator bias through a baseline analysis of annotator perceptions of meaning. Second, the final labels by annotators undergo reconciliation and validation by domain experts living in Chicago neighborhoods with high rates of violence. Annotation is provided by trained social work student annotators and domain experts, community members who live in neighborhoods from which the Twitter data derives. Social work students are rigorously trained in textual and discourse analysis methods using the adapted and modified DUVAA method described above. Our domain experts consist of Black and Latino men and women who affiliate with Chicago-based violence prevention programs. While our domain experts leverage their community expertise to annotate the Twitter data, our social work annotators undergo a five stage training process to prepare them for eliciting context and nuance from the corpus.

For annotation we used the open-source annotation platform VATAS (Patton et al. 2019b) with the following annotation tasks:

Concepts/Codes	Twitter	Tumblr	Total
<i>handgun</i>	164	41	205
<i>long gun</i>	15	105	116
<i>joint</i>	185	113	298
<i>marijuana</i>	56	154	210
<i>person</i>	1368	74	1442
<i>tattoo</i>	227	33	260
<i>hand gesture</i>	572	2	574
<i>lean</i>	43	116	159
<i>money</i>	107	138	245
<i>aggression</i>	457 (185)	-	457 (185)
<i>loss</i>	397 (308)	-	397 (308)
<i>substance use</i>	365 (268)	-	365 (268)

Table 1: Numbers of instances for the different visual concepts and psychosocial codes in our dataset. For the different codes, the first number indicates for how many tweets at least one annotator assigned the corresponding code, numbers in parentheses are based on per-tweet majority votes.

- In the *bounding box annotation task*, annotators are shown the text and tweet of the image. Annotators are asked to mark all local visual concepts of interest by drawing bounding boxes directly on the image. For each image we collected two annotations.
- To reconcile all conflicts between annotations we implemented a *bounding box reconciliation task* where conflicting annotations are shown side by side and the better annotation can be chosen by the third annotator.
- For *code annotation*, tweets including the text, image and link to the original post, are displayed and for each of the three codes *aggression*, *loss* and *substance use*, there is a checkbox the annotator is asked to check if the respective code applies to the tweet. We collected two student annotations and two domain expert annotations for each tweet. In addition, we created one extra code annotation to break ties for all tweets with any disagreement between the student annotations.

Our social work colleagues took several measures to ensure the quality of the resulting dataset during the annotation process. Annotators met weekly as a group with an expert annotator to address any challenges and answer any questions that came up that week. This process also involved iterative correction of reoccurring annotation mistakes and infusion of new community insights provided by domain experts. Before the meeting each week, the expert annotator closely reviewed each annotator’s interpretations and labels to check for inaccuracies.

During the annotation process, we monitored statistics of the annotated concepts. We were aiming for at least around 100-200 instances for training plus additional instances for testing, and preliminary statistics made us realize that for some visual concepts of interest, the number of expected instances in the final dataset was insufficient. Specifically, this affected the concepts *handgun*, *long gun*, *money*, *marijuana*, *joint*, and *lean*. For all of these concepts we crawled additional images from Tumblr, using the public Tumblr

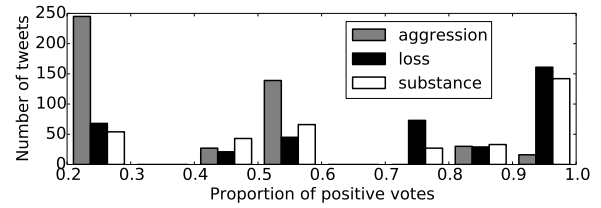


Figure 3: Annotator consensus for all psychosocial codes. For better visibility, we exclude tweets that were unanimously annotated as not belonging to the respective codes. Note that for each tweet there are 4 or 5 code annotations.

API with a keyword-based approach for the initial crawling. We then manually filtered the images we retrieved to obtain around 100 images for each of these specific concepts. Finally we put these images into our annotation system and annotated them w.r.t. all local visual concepts listed in Section 3.2.

3.5 Statistics

The distribution of concepts in our dataset is shown in Table 1. Note that in order to ensure sufficient quality of the annotations, but also due to the nature of the data, we relied on a special annotation process and kept the total size of the dataset comparatively small.

Figure 3 displays the distributions of fractions of positive votes for all 3 psychosocial codes. These statistics indicate that for the code *aggression*, disagreement between annotators is substantially higher than for the codes *loss* and *substance use*, which both display a similar pattern of rather high annotator consensus.

3.6 Data Sharing and Ethical Considerations

The users in our dataset comprise youth of color from marginalized communities in Chicago with high rates of gun violence. Careless handling of the data has the potential to further marginalize and harm the users who are already vulnerable to surveillance and criminalization by law enforcement. Thus, we take several special precautions to protect these users. This includes only sharing the data with people who sign our ethical guidelines, and only releasing tweet IDs instead of any actual tweet contents.³

Our social work team members initially attempted to seek informed consent, but to no avail, as participants did not respond to requests. To protect users, we altered text during any presentation so that tweets are not searchable on the Internet, excluded all users that were initially private or changed their status to private during the analysis, and consulted Chicago-based domain experts on annotation decisions, labels and dissemination of research. More general

³We will make tweet IDs for the data available to researchers who sign an MOU specifying their intended use of the data and their agreement with our ethical guidelines. Contact Philipp Blandfort (philipp.blandfort@dfki.de) or Shih-Fu Chang (sc250@columbia.edu). Our code is available at <https://gitlab.com/blandfort/multimodal>.

ethical implications of this study will be addressed in Section 6.3.

4 Methods for Multimodal Analysis

In this section we describe the building blocks for analysis, the text features and image features used as input for the psychosocial code classification with an SVM, and the multimodal fusion methods we explored. Details of implementation and analysis of results will be presented in Sections 5 and 6.

4.1 Text Features

As text features, we exploit both sparse linguistic features as well as dense vector representations extracted from a CNN classifier operating at either the word or character level.

Linguistic Features To obtain the linguistic features, we used the feature extraction code of (Blevins et al. 2016) from which we obtained the following:

- Unigram and bigram features.
- Part-of-Speech (POS) tagged unigram and bigram features. The POS tagger used to extract these features was adapted to this domain and cohort of users.
- The minimum and maximum pleasantness, activation, and imagery scores of the words in the input text. These scores are computed by looking up each word’s associated scores in the Dictionary of Affective Language (DAL). Vernacular words and emojis were mapped to the Standard American English of the DAL using a translation phrasebook derived from this domain and cohort of users.

CNN Features To extract the CNN features we train binary classifiers for each code. We use the same architecture for both the word and character level models and so we describe only the word level model below. Our CNN architecture is roughly the same as (Kim 2014) but with an extra fully connected layer before the final softmax. I.e., the text is represented as a sequence of embeddings, over which we run a series of varying width one-dimensional convolutions with max-pooling and a pointwise-nonlinearity; the resultant convolutional feature maps are concatenated and fed into a multi-layer perceptron (MLP) with one hidden layer and softmax output. After training the network, the softmax layer is discarded, and we take the hidden layer output in the MLP as the word or character feature vector to train the psychosocial code SVM.

4.2 Image Features

We here describe how we extract visual features from the images that will be fed to the psychosocial code classifier.

Local Visual Concepts To detect the local concepts defined in section 3.2, we adopt the Faster R-CNN model (Ren et al. 2017), a state-of-the-art method for object detection in images. The Faster R-CNN model introduced a *Region Proposal Network* (RPN) to produce region bounds and objectness score at each location of a regular grid. The bounding boxes proposed by the RPN are fed to a Fast R-CNN (Girshick 2015) detection network. The two networks share their

convolutional features, enabling the whole Faster R-CNN model to be trained end-to-end and to produce fast yet accurate detections. Faster R-CNN has been shown (Huang et al. 2017) to be one of the best models among the modern convolutional object detectors in terms of accuracy. Details on the training of the model on our data are provided in Section 5.2. We explore the usefulness of the local visual concepts in two ways:

- For each *local visual concept* detected by the faster R-CNN, we count the frequency of the concept detected in a given image. For this, we only consider predictions of the local concept detector with a confidence higher than a given threshold, which is varied in experiments.
- In order to get a better idea of the potential usefulness of our proposed local visual concepts, we add one model to the experiments that uses *ground truth local concepts* as features. This corresponds to features from a perfect local visual concept detector. This method is considered out-of-competition and is not used for any fusion methods. It is used only to gain a deeper understanding of the relationship between the local visual concepts and the psychosocial codes.

Global Features As *global image features* we process the given images using a deep convolutional model (Inception-v3 (Szegedy et al. 2016)) pre-trained on ImageNet (Deng et al. 2009) and use activations of the last layer before the classification layer as features. We decided not to update any weights of the network due to the limited size of our dataset and because such generic features have been shown to have a strong discriminative power (Razavian et al. 2014).

4.3 Fusion Methods for Code Detection

In addition to the text- and image-only models that can be obtained by using individually each feature described in Sections 4.1 and 4.2, we evaluate several tweet classification models that combine multiple kinds of features from either one or both modalities. These approaches always use features of all non-fusion methods for the respective modalities outlined in Sections 4.1 and 4.2, and combine information in one of the following two ways:

- *Early fusion*: the different kinds of features are concatenated into a single feature vector, which is then fed into the SVM. For example, the text-only early fusion model first extracts linguistic features and deploys a character and a word level CNN to compute two 100-dimensional representations of the text, and then feeds the concatenation of these three vectors into an SVM for classification.
- *Late fusion* corresponds to an ensemble approach. Here, we first train separate SVMs on the code classification task for each feature as input, and then train another final SVM to detect the psychosocial codes from the probabilistic outputs of the previous SVMs.

5 Experiments

Dividing by twitter users (splitting on a user basis so that tweets of the same user are not repeated in both training and

test sets), we randomly split our dataset into 5 parts with similar code distributions and total numbers of tweets. We use these splits for 5-fold cross validation, i.e. all feature representations that can be trained and the psychosocial code prediction models are trained on 4 folds and tested on the unseen 5th fold. All reported performances and sensitivities are averaged across these 5 data splits. Statements on statistical significance are based on 95% confidence intervals computed from the 5 values on the 5 splits.

We first detail how the text and image representations are trained on our data. We then discuss the performance of different uni- and multimodal psychosocial code classifiers. The last two experiments are designed to provide additional insights into the nature of the code classification task and the usefulness of specific concepts.

5.1 Learning Text Representations

Linguistic Features We do not use all the linguistic features described in Section 4.1 as input for the SVM but instead during training apply feature selection using an ANOVA F-test that selects the top 1,300 most important features. Only the selected features are provided to the SVM for classification. We used the default SVM hyperparameter settings of (Blevins et al. 2016).

CNN Features We initialize the word embeddings with pretrained 300-dimensional *word2vec* (Mikolov et al. 2013) embeddings (obtained from <https://code.google.com/p/word2vec/>). For the character level model, we used 100-dimensional character embeddings randomly initialized by sampling uniformly from $(-0.25, 0.25)$. In both CNN models we used convolutional filter windows of size 1 to 5 with 100 feature maps each. The convolutional filters applied in this way can be thought of as word (or character) ngram feature detectors, making our models sensitive to chunks of one to five words (or characters) long. We use a 100-dimensional hidden layer in the MLP. During cross-validation we train the CNNs using the Nesterov Adam (Dozat 2016) optimizer with a learning rate of .002, early stopping on 10% of the training fold, and dropout of .5 applied to the embeddings and convolutional feature maps.

5.2 Learning to Detect Local Concepts

Our local concepts detector is trained using the image data from Twitter and Tumblr and the corresponding bounding box annotations. We use the Twitter data splits defined above and similarly define five splits for the Tumblr data with similar distribution of concepts across different parts. We train a Faster R-CNN model (publicly available implementation from <https://github.com/endernewton/tf-faster-rcnn>) using a 5-fold cross validation, training using 4 splits of the Twitter and Tumblr data joined as a training set. We evaluate our local concepts detection model on the joined test set, as well as separately on the Twitter and Tumblr test set, and will discuss its performance in section 6.1.

The detector follows the network architecture of VGG-16 and is trained using the 4-step alternating training approach detailed in (Ren et al. 2017). The network is initialized with an ImageNet-pretrained model and trained for the task of

local concepts detection. We use an initial learning rate of 0.001 which is reduced by a factor of 0.9 every 30k iterations and trained the model for a total of 250k iterations. We use a momentum of 0.8 and a weight decay of 0.001.

During training, we augment the data by flipping images horizontally. In order to deal with class imbalance while training, we weigh the classification cross entropy loss for each class by the logarithm of the inverse of its proportion in the training data. We will discuss in detail the performance of our detector in Section 6.1.

5.3 Detecting Psychosocial Codes

We detect the three psychosocial codes separately, i.e. for each code we consider the binary classification task of deciding whether the code applies to a given tweet.

For our experiments we consider a tweet to belong to the positive class of a certain code if at least one annotator marked the tweet as displaying that code. For the negative class we used all tweets that were not marked by any annotator as belonging to the code (but might belong or not belong to any of the two other codes). We chose this way of converting multiple annotations to single binary labels because our final system is not meant to be used as a fully automatic detector but as a pre-filtering mechanism for tweets that are potentially useful for social workers. Given that the task of rating tweets with respect to such psychosocial codes inevitably depends on the perspective on the annotator to a certain extent, we think that even in case of a majority voting mechanism, important tweets might be missed.⁴

In addition to the models trained using the features described in Section 4, we also evaluate two baselines that do not process the actual tweet data in any way. Our *random baseline* uses the training data to calculate the prior probability of a sample belonging to the positive class and for each test sample predicts the positive class with this probability without using any information about the sample itself. The other baseline, *positive baseline*, always outputs the positive class.

All features except the linguistic features were fed to an SVM using the RBF kernel for classifying the psychosocial codes. For linguistic features, due to issues when training with an RBF kernel, we used a linear SVM with squared hinge loss, as in (Blevins et al. 2016), and $C = 0.01, 0.03$ and 0.003 for detecting *aggression*, *loss* and *substance use* respectively. Class weight was set to balanced, with all other parameters kept at their default values. We used the SVM implementation of the Python library scikit-learn (Pedregosa et al. 2011). This two stage approach of feature extraction plus classifier was chosen to allow for a better understanding of the contributions of each feature. We preferred SVMs in the 2nd stage over deep learning methods since SVMs can be trained on comparatively smaller datasets without the need to optimize many hyperparameters.

⁴For future work we are planning to have a closer look at the differences between annotations of community experts and students and based on that treat these types of annotations differently. We report a preliminary analysis in that direction in Section 6.2.

Modality	Features	Fusion	Aggression				Loss				Substance Use				mAP
			P	R	F1	AP	P	R	F1	AP	P	R	F1	AP	
-	-(random baseline)	-	0.25	0.26	0.26	0.26	0.17	0.17	0.17	0.20	0.18	0.18	0.18	0.20	0.23
-	-(positive baseline)	-	0.25	1.00	0.40	0.25	0.21	1.00	0.35	0.22	0.20	1.00	0.33	0.20	0.22
text	linguistic features	-	0.35	0.34	0.34	0.31	0.71	0.47	0.56	0.51	0.25	0.53	0.34	0.24	0.35
text	CNN-char	-	0.37	0.47	0.39	0.36	0.75	0.66	0.70	0.77	0.27	0.32	0.29	0.28	0.45
text	CNN-word	-	0.39	0.46	0.42	0.41	0.71	0.65	0.68	0.77	0.28	0.30	0.29	0.31	0.50
text	all textual	early	0.40	0.46	0.43	0.42	0.70	0.73	0.71	0.81	0.25	0.37	0.30	0.30	0.51
text	all textual	late	0.43	0.41	0.42	0.42	0.69	0.65	0.67	0.79	0.29	0.37	0.32	0.32	0.51
image	inception global	-	0.43	0.64	0.51	0.49	0.38	0.57	0.45	0.43	0.41	0.62	0.49	0.48	0.47
image	Faster R-CNN local (0.1)	-	0.43	0.64	0.52	0.47	0.28	0.56	0.37	0.31	0.44	0.30	0.35	0.37	0.38
image	Faster R-CNN local (0.5)	-	0.47	0.48	0.47	0.44	0.30	0.39	0.33	0.31	0.46	0.12	0.19	0.30	0.35
image	all visual	early	0.49	0.62	0.55	0.55*	0.38	0.57	0.45	0.44	0.41	0.59	0.48	0.48	0.49
image	all visual	late	0.48	0.51	0.49	0.52	0.40	0.51	0.44	0.43	0.47	0.52	0.50	0.51*	0.49
image+text	all textual + visual	early	0.48	0.51	0.49	0.53	0.72	0.73	0.73	0.82*	0.37	0.53	0.43	0.45	0.60
image+text	all textual + visual	late	0.48	0.44	0.46	0.53	0.71	0.67	0.69	0.80	0.44	0.43	0.43	0.48	0.60*

Table 2: Results for detecting the psychosocial codes: aggression, loss and substance use. For each code we report precision (P), recall (R), F1-scores (F1) and average precision (AP). The last column describes overall performance in terms of mean average precision (mAP) across all three codes. Numbers shown are mean values of 5-fold cross validation performances. The highest performance (based on AP) for each code is marked with an asterisk. In bold we highlight all performances not significantly worse than the highest one (based on statistical testing with 95% confidence intervals).

For all models we report results with respect to the following metrics: precision, recall and F1-score (always on positive class), and average precision (using detector scores to rank output). The former 3 measures are useful to form an intuitive understanding of the performances, but for drawing all major conclusions we rely on average precision, which is an approximation of the area under the entire precision-recall curve, as compared to measurement at only one point.

The results of our experiments are shown in Table 2. Our results indicate that image and text features play different roles in detecting different psychosocial codes. Textual information clearly dominates the detection of code *loss*. We hypothesize that loss is better conveyed textually whereas substance use and aggression are easier to express visually. Qualitatively, the linguistic features with the highest magnitude weights (averaged over all training splits) in a linear SVM bear this out, with the top five features for loss being i) *free*, ii) *miss*, iii) *bro*, iv) *love* v) *you*; the top five features for substance use being i) *smoke*, ii) *cup*, iii) *drank*, iv) *@mention* v) *purple*; and the top five features for aggression being i) *Middle Finger Emoji*, ii) *Syringe Emoji*, iii) *opps*, iv) *pipe* v) *2017*. The loss features are obviously related to the death or incarceration of a loved one (e.g. *miss* and *free* are often used in phrases wishing someone was freed from prison). The top features for aggression and substance use are either emojis which are themselves pictographic representations, i.e. not a purely textual expression of the code, or words that reference physical objects (e.g. *pipe*, *smoke*, *cup*) which are relatively easy to picture.

Image information dominates classification of both the *aggression* and *substance use* codes. Global image features tend to outperform local concept features, but combining local concept features with global image features achieves the best image-based code classification performance. Importantly, by fusing both image and text features, the combined detector performs consistently very well for all three codes, with the mean average precision (mAP) across the three codes being 0.60, compared to 0.51 for the text only

detector and 0.49 for the image only detector. This demonstrates a relative gain in mAP of around 20% of the multimodal approach over any single modality.

5.4 Sensitivity Analysis

We performed additional experiments to get a better understanding of the usefulness of our local visual concepts for the code prediction task. For sensitivity analysis we trained linear SVMs on psychosocial code classification, using as features either the local visual concepts detected by Faster R-CNN or the ground truth visual concepts. In general, the sensitivity score of any input feature x is calculated as the partial derivative of the model’s output with respect to x and thus quantifies how changes in x affect the model’s decision. In our case, these partial derivatives correspond to the coefficients of the linear SVM (due to linearity of the model). All reported sensitivity scores are average values of the corresponding coefficients of the linear SVM, computed across the 5 folds used for the code detection experiments. Results from this experiment can be found in Table 3.

From classification using ground truth visual features we see that for detecting *aggression*, the local visual concepts *handgun* and *long gun* are important, while for detecting *substance use*, the concepts *marijuana*, *lean*, *joint* are most significant. For the code *loss*, *marijuana* as the most relevant visual concept correlates negatively with *loss*, but overall, significance scores are much lower.

Interestingly, the model that uses the higher detection score threshold of 0.5 for the local visual concept detection behaves similarly to the model using ground truth annotations, even though the classification performance is better with the lower threshold. This could indicate that using a lower threshold makes the code classifier learn to exploit false alarms of the concept detector.

However, it needs to be mentioned that sensitivity analysis can only measure how much the respective classifier relies on the different parts of the input, given the respective overall setting. This can provide useful information about which

Concept	Aggression			Loss			Substance Use		
	0.1	0.5	GT	0.1	0.5	GT	0.1	0.5	GT
<i>handgun</i>	0.73	0.93	1.05	0.06	0.10	0.06	0.06	0.09	0.11
<i>long gun</i>	0.26	0.91	1.30	-0.17	0.14	0.14	0.42	0.04	-0.47
<i>joint</i>	0.42	-0.08	0.05	-0.15	0.00	0.10	0.25	1.3	1.41
<i>marijuana</i>	0.17	0.18	0.12	-0.19	-0.45	-0.35	0.93	1.29	1.47
<i>person</i>	0.34	-0.01	-0.17	0.11	0.10	0.12	0.04	0.28	-0.01
<i>tattoo</i>	-0.11	-0.09	0.01	-0.02	0.03	-0.03	0.04	0.06	-0.02
<i>hand gesture</i>	0.20	0.67	0.53	-0.01	0.12	0.05	0.01	0.06	-0.02
<i>lean</i>	-0.07	0.03	-0.28	-0.20	-0.06	-0.14	0.68	0.59	1.46
<i>money</i>	-0.06	0.06	-0.02	0.00	-0.01	-0.01	0.18	-0.04	-0.19
F1	0.51	0.46	0.65	0.37	0.33	0.38	0.34	0.17	0.76
AP	0.41	0.39	0.54	0.29	0.28	0.30	0.33	0.27	0.72

Table 3: Sensitivity of visual local concept based classifiers with respect to the different concepts. For each of the three psychosocial codes, we include two versions that use detected local concepts (“0.1” and “0.5”, where the number indicates the detection score threshold) and one version that uses local concept annotations as input (“GT”).

parts are *sufficient* for obtaining comparable detection results, but there is no guarantee that the respective parts are also *necessary* for achieving the same classification performance. For example, imagine that two hypothetical concepts A and B correlate perfectly with a given class and a detector for this class is given both concepts as input. The detector could make its decision based on A alone, but A is not really necessary since the same could be achieved by using B instead. For this reason, we ran an ablation study to get quantitative measurements on the necessity of local visual concepts for code classification.

5.5 Ablation Study

In our ablation study we repeated the psychosocial code classification experiment using ground truth local visual concepts as features, excluding one concept at a time to check how this affects overall performance of the model.

We found that for *aggression*, removing the concepts *handgun* or *hand gesture* leads to the biggest drops in performance, while for *substance use*, the concepts *joint*, *marijuana* and *lean* are most important. For *loss*, removal of none of the concepts causes any significant change. See Table 4 for further details.

6 Open Challenges

In this section, we provide a more in-depth analysis of what makes our problem especially challenging and how we plan to address those challenges in the future.

6.1 Local Concepts Analysis

We report in Table 5 the average precision results of our local concept detection approach on the “Complete” test set, i.e. joining data from both Twitter and Tumblr, and separately on the Twitter and Tumblr test sets. We compute the average precision on each test fold separately and report the average and standard deviation values over the 5 folds. When looking at the results on the “Complete” test set, we see average precision values ranging from 0.26 on *tattoo* to

Removed Concept	Aggression		Substance Use	
	F1	AP	F1	AP
<i>handgun</i>	-0.10	-0.15	-0.01	0.01
<i>long gun</i>	-0.01	-0.01	-0.00	-0.00
<i>joint</i>	0.00	-0.00	-0.35	-0.28
<i>marijuana</i>	0.00	0.00	-0.09	-0.09
<i>person</i>	-0.01	-0.01	-0.01	-0.00
<i>tattoo</i>	0.00	0.00	0.01	-0.00
<i>hand gesture</i>	-0.13	-0.09	0.00	0.00
<i>lean</i>	-0.00	0.00	-0.07	-0.07
<i>money</i>	0.00	0.00	0.00	0.00

Table 4: Differences in psychosocial code detection performance of detectors with specific local concepts removed as compared to a detector that uses all local concept annotations. (Numbers less than 0 indicate that removing the concept reduces the corresponding score.) Bold font indicates that the respective number is significantly less than 0. For the code loss none of the numbers was significantly different from 0, hence we decided to not list them in this table.

0.80 for *person* and the mean average precision of 0.54 indicating a rather good performance. This results on the “Complete” test set hides two different stories, however, as the performance is much lower on the Twitter test set (mAP of 0.29) than on the Tumblr one (mAP of 0.81).

As detailed in Section 3.4, we have crawled additional images, especially targeting the concepts with a low occurrence count in Twitter data as detailed in Table 1. However, crawling images from Tumblr targeting keywords related to those concepts lead us to gather images where the target concept is the main subject in the image, while in our Twitter images they appear in the image but are rarely the main element in the picture. Further manually analyzing the images crawled from Twitter and Tumblr, we have confirmed this “domain gap” between the two sources of data that can explain the difference of performance. This puts in light the challenges associated with detecting these concepts in our Twitter data. We believe the only solution is therefore to gather additional

Concept	Complete	Twitter	Tumblr
	AP \pm SD	AP \pm SD	AP \pm SD
<i>handgun</i>	0.30 \pm 0.07	0.13 \pm 0.02	0.74 \pm 0.11
<i>long gun</i>	0.78 \pm 0.03	0.29 \pm 0.41	0.85 \pm 0.05
<i>joint</i>	0.30 \pm 0.07	0.01 \pm 0.01	0.57 \pm 0.04
<i>marijuana</i>	0.73 \pm 0.08	0.28 \pm 0.17	0.87 \pm 0.09
<i>person</i>	0.80 \pm 0.03	0.80 \pm 0.03	0.95 \pm 0.03
<i>tattoo</i>	0.26 \pm 0.06	0.08 \pm 0.02	0.84 \pm 0.06
<i>hand gesture</i>	0.27 \pm 0.05	0.28 \pm 0.04	0.83 \pm 0.29
<i>lean</i>	0.78 \pm 0.07	0.38 \pm 0.15	0.87 \pm 0.03
<i>money</i>	0.60 \pm 0.02	0.35 \pm 0.08	0.73 \pm 0.05
mAP	0.54 \pm 0.01	0.29 \pm 0.05	0.81 \pm 0.02

Table 5: Local concepts detection performance.

images from Twitter from similar users. This will be part of the future work of this research.

The local concepts are highly relevant for the detection of the codes *aggression* and *substance use* as it can be highlighted in the column GT in Table 3 and from the ablation study reported in Table 4. The aforementioned analysis of the local concepts detection limitation on the Twitter data explains why the performance using the detected concepts is substantially lower than when using ground truth local concepts. We will therefore continue to work on local concepts detection in the future as we see they could provide significant help in detecting these two codes and also because they would help in providing a clear interpretability of our model.

6.2 Annotation Analysis

In order to identify factors that led to divergent classification between social work annotators and domain experts, we reviewed 10% of disagreed-upon tweets with domain experts. In general, knowledge of local people, places, and behaviors accounted for the majority of disagreements (Patton et al. 2019a). In particular, recognizing and having knowledge of someone in the image (including their reputation, gang affiliation, and whether or not they had been killed or incarcerated) was the most common reason for disagreement between our annotators and domain experts. Less commonly, identifying or recognizing physical items or locations related to the specific cultural context of the Chicago area (e.g., a home known to be used in the sale of drugs) also contributed to disagreement. The domain experts’ nuanced understanding of hand signs also led to a more refined understanding of the images, which variably increased or decreased the perceived level of aggression. For example, knowledge that a certain hand sign is used to disrespect a specific gang often resulted in increased perceived level of aggression. In contrast, certain hand gestures considered to be disrespectful by our social work student annotators (e.g., displaying upturned middle fingers) were perceived to be neutral by domain experts and therefore not aggressive. Therefore, continuous exchange with the domain experts is needed to always ensure that the computer scientists are aware of all these aspects when further developing their methods.

6.3 Ethical Implications

Our team was approached by violence outreach workers in Chicago to begin to create a computational system that would enhance violence prevention and intervention. Accordingly, our automatic vision and textual detection tools were created to assist social workers in their efforts to understand and prevent community violence through social media, but not to optimize any systems of surveillance. This shift away from identifying potentially violent users to understanding pathways to violent online content highlights systemic gaps in economic, educational, and health-related resources that are often root causes to violent behavior. Our efforts for ethical and just treatment of the users who provide our data include removal of identifying information during presentation of work (e.g., altering text to eliminate searchability), the inclusion of Chicago-based community members as domain experts in the analysis and validation of our findings, only sharing the data with researchers who sign our ethical guidelines, and only releasing tweet IDs instead of actual tweet contents. Our long term efforts include using multimodal analysis to enhance current violence prevention efforts by providing insight into social media behaviors that may shape future physical altercations.

7 Conclusion

We have introduced the problem of multimodal social media analysis for gang violence prevention and presented a number of automatic detection experiments to gain insights into the expression of *aggression*, *loss* and *substance use* in tweets coming from this specific community, measure the performance of state-of-the-art methods on detecting these codes in tweets that include images, and analyze the role of the two modalities text and image in this multimodal tweet classification setting.

We proposed a list of general-purpose local visual concepts and showed that despite insufficient performance of current local concept detection, when combined with global visual features, these concepts can help visual detection of *aggression* and *substance use* in tweets. In this context we also analyzed in-depth the contribution of all individual concepts.

In general, we found the relevance of the text and image modalities in tweet classification to depend heavily on the specific code being detected, and demonstrated that combining both modalities leads to a significant improvement of overall performance across all 3 psychosocial codes.

Findings from our experiments affirm prior social science research indicating that youth use social media to respond to, cope with, and discuss their exposure to violence. Human annotation, however, remains an important element in vision detection in order to understand the culture, context and nuance embedded in each image. Hence, despite promising detection results, we argue that psychosocial code classification is far from being solved by automatic methods. Here our interdisciplinary approach clearly helped to become aware of the whole complexity of the task, but also to see the broader context of our work, including important ethical implications which were discussed above.

Acknowledgments

During some of this work the first author was staying at Columbia University. This research stay was supported by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD). Furthermore, we would like to thank the reviewers for their valuable feedback. Last but not least, we thank all our annotators: Allison Aguilar, Rebecca Carlson, Natalie Hession, Chloe Martin, Mirinda Morency.

References

- Atkinson, R., and Flint, J. 2001. Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social Research Update* 33.
- BBC World News. 2017. *Click, Pre Crime, Chicago's crime-predicting software*. <http://www.bbc.co.uk/programmes/p052fll7>.
- Blevins, T.; Kwiatkowski, R.; Macbeth, J.; McKeown, K.; Patton, D.; and Rambow, O. 2016. Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2196–2206.
- Citizens Crime Commission of New York City. 2017a. *E-Responder: a brief about preventing real world violence using digital intervention*. www.nycrimecommission.org/pdfs/e-responder-brief-1.pdf.
- City of Chicago. 2017b. *Strategic Subject List*. <https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np>.
- Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, 379–387.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Li, F.-F. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE Computer Society.
- Dozat, T. 2016. Incorporating nesterov momentum into adam. Technical report, Stanford University.
- Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88(2):303–338.
- Gerber, M. S. 2014. Predicting crime using twitter and kernel density estimation. *Decision Support Systems* 61:115 – 125.
- Girshick, R. 2015. Fast r-cnn. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, 1440–1448. IEEE.
- Hanna, A. 2017. Mpedts: Automating the generation of protest event data. *Deposited at SocArXiv* <https://osf.io/preprints/socarxiv/xuqmv>.
- Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*.
- Kapustin, M.; Ludwig, J.; Punkay, M.; Smith, K.; Spiegel, L.; and Welgus, D. 2017. Gun violence in chicago, 2016. *Chicago, IL: University of Chicago Crime Lab*.
- Keith, K. A.; Handler, A.; Pinkham, M.; Magliozzi, C.; McDuffie, J.; and O'Connor, B. 2017. Identifying civilians killed by police with distantly supervised entity-event extraction. *arXiv preprint arXiv:1707.07086*.
- Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Lane, J. 2016. The digital street: An ethnographic study of networked street life in harlem. *American Behavioral Scientist* 60(1):43–58.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13*, 3111–3119. USA: Curran Associates Inc.
- Nellis, A.; Greene, J.; Mauer, M.; and (U.S.), S. P. 2008. *Reducing Racial Disparity in the Criminal Justice System: A Manual for Practitioners and Policymakers*. Sentencing Project.
- Patton, D. U.; McKeown, K.; Rambow, O.; and Macbeth, J. 2016. Using natural language processing and qualitative analysis to intervene in gang violence: A collaboration between social work researchers and data scientists. *arXiv preprint arXiv:1609.08779*.
- Patton, D. U.; Lane, J.; Leonard, P.; Macbeth, J.; and Smith Lee, J. R. 2017. Gang violence on the digital street: Case study of a south side chicago gang member's twitter communication. *new media & society* 19(7):1000–1018.
- Patton, D.; Blandfort, P.; Frey, W.; Gaskell, M.; and Karaman, S. 2019a. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
- Patton, D.; Blandfort, P.; Frey, W. R.; Gaskell, M.; Schifanella, R.; and Chang, S.-F. 2019b. VATAS: An open-source web platform for visual and textual analysis of social media. *Journal of the Society for Social Work and Research*. In press.
- Patton, D. U.; Eschmann, R. D.; and Butler, D. A. 2013. Internet banging: New trends in social media, gang violence, masculinity and hip hop. *Computers in Human Behavior* 29(5):A54–A59.
- Pavlick, E.; Ji, H.; Pan, X.; and Callison-Burch, C. 2016. The gun violence database: A new task and data set for nlp.

In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1018–1024.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Razavian, A. S.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, 512–519. IEEE.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39(6):1137–1149.

Schmidt, C. 2018. *Holding algorithms (and the people behind them) accountable is still tricky, but doable*. Nieman Lab. <http://nie.mn/2ucDuw2>.

Sheley, K. 2017. *Statement on Predictive Policing in Chicago*. ACLU of Illinois. <http://www.aclu-il.org/en/press-releases/statement-predictive-policing-chicago>.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2818–2826.

VATAS: An Open-Source Web Platform for Visual and Textual Analysis of Social Media

Desmond U. Patton *Columbia University*

Philipp Blandfort *TU Kaiserslautern and German Research Center for Artificial Intelligence*

William R. Frey *Columbia University*

Rossano Schifanella *University of Turin*

Kyle McGregor *NYU Langone Health*

Shih-Fu U. Chang *Columbia University*

ABSTRACT Social media have created a new environmental context for the study of social and human behavior and services. Although social work researchers have become increasingly interested in the use of social media to address social problems, they have been slow to adapt tools that are flexible and convenient for analyzing social media data. They have also given inadequate attention to bias and representation inherent in many multimedia data sets. This article introduces the Visual and Textual Analysis of Social Media (VATAS) system, an open-source Web-based platform for labeling or annotating social media data. We use a case study approach, applying VATAS to a study of Chicago, IL, gang-involved youth communication on Twitter to highlight VATAS' features and opportunities for interdisciplinary collaboration. VATAS is highly customizable, can be privately held on a secure server, and allows for export directly into a CSV file for qualitative, quantitative, and machine-learning analysis. Implications for research using social media sources are noted.

KEYWORDS: social media, data analysis, qualitative, machine learning, multimedia

doi: 10.1086/707667

Social work researchers seek rigorous and innovative ways to study social phenomena and the complexities of human behavior, population health, and broader social problems. Findings from traditional qualitative and quantitative methods have shaped social policy, practice, and research in the areas of health, mental health, and social inequality. In recent years, the American Academy of Social Work and Social Welfare (2018) has adopted the use of technology for research and practice as a grand challenge, recognizing that the field demands innovation in order to “accelerate the pace of social discovery” (p. 1, para. 2). Although the types, quantity, and availability of data have multiplied in recent years (Kitchin, 2014)—social work research is no longer limited to time consuming and expensive data collection

000 Journal of the Society for Social Work & Research Spring 2020

methods such as surveys, chart reviews, and in-person interviews—the types of data used to answer social work research questions have been slow to change.

Social media are useful yet underutilized data sources that can be valuable to social work research. Social media (also known as social networking sites) “employ mobile and Web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss, and modify user-generated content” (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011; p. 241). Three core features define social media: (a) user-generated content and sharing; (b) user-created profiles that are maintained by the platform; and (c) online social networks created by connecting users with other individuals or groups (Kietzmann et al., 2011). Popular social media sites include but are not limited to Facebook, Twitter, Instagram, Tumblr, Reddit, wechat, and Snapchat. Social media generate several types of data, including user-generated content or “posts,” user profiles, and relationships between users. An example of user-generated content is a “tweet,” which is a text-based Twitter comment of less than 280 characters created by a user on almost any topic. The collection of social media data, with or without additional metadata (e.g., source, time, or annotations), form a social media data set. (See the online Appendix for a glossary of terms.) One common type of social media data set is a social media corpus, which is defined as set of social media posts (with or without metadata) collected using a systematic methodology. Social media data sets are growing in popularity within social work and social science research communities as access to billions of people has transformed the amount and type of social data researchers can collect and analyze. Many forms of user-generated social media data are posted voluntarily and unprompted, which may offer different insights into people’s lives than other quantitative and qualitative research and data collection methods.

Social media data sets have primarily been collected and analyzed by researchers in computer science and data science, where the availability of large amounts of social media data has advanced approaches in machine learning (Thomee et al., 2016). Computer scientists have developed efficient Web-based tools for collecting additional information (e.g., labels, bounding boxes) that can be used to capture individual and group behaviors (Russell, Torralba, Murphy, & Freeman, 2008). In-house annotators may label social media, or the task may be outsourced to crowdsourcing websites such as Amazon Mechanical Turk, where registered users (who are unknown to the researchers conducting the data collection) are paid to perform annotations. Although these systems have been quite useful for annotating large amounts of social media data, they usually require some form of payment and are not designed for complex analysis of individual items. In addition, identifying crowdworkers with the required understanding of context in language use, emoji, images, and music would be nearly impossible in our most marginalized and systemically underresourced communities, where much social work and social science research takes place. Thus, although the use of social media has great promise

in enhancing social work research by providing new data points on topics such as mental health, trauma, and interpersonal violence, analyzing millions of social media posts without considering the context and culture embedded in posts could lead to dangerous assumptions that have severe consequences for marginalized communities (Frey, Patton, Gaskell, & McGregor, 2018). Although social workers are trained to be sensitive to such contextual factors, social work researchers have been slow to adapt flexible, convenient tools for annotating social media data (Coulton, Goerge, Putnam-Hornstein, & de Haan, 2015), especially given the need to address bias and representation in multimedia data sets (Blandfort et al., 2019).

In this paper, we describe the development of the Visual and Textual Analysis of Social Media (VATAS) tool, a flexible, Web-based solution to analyze and annotate social media data. We share code for its development and deployment and demonstrate how VATAS can be adopted for rigorous scientific study of social media corpora, particularly from marginalized communities. VATAS is free and designed to work in tandem with machine-learning approaches, allowing researchers to select and train annotators based on their needs (e.g., extracting culture, nuance, and local context) and to use annotated posts for training machine-learning models or running statistical analyses. We explain how to organize such a collaboration and believe that, overall, our study makes a good case for strengthening interdisciplinary research, especially in the context of annotating social media data.

Methods for Collecting and Annotating Social Media Data

The overall goal is to annotate social media data that include images and require domain expertise (e.g., local language or subculture) or a specific educational background (e.g., social work or social science), and to do so in such a way that qualitative analysis directly feeds into annotation data for automatic processing by computer scientists. This concurrent process relies on collecting specific types of annotations, such as bounding boxes or multiple-choice answers, that can be exported in a format that is easy to process automatically (e.g., JSON or CSV). In this section, we outline methods for collecting and annotating social media data. These methods include machine learning, crowdsourcing, open-source projects for crowdsourcing, qualitative data analysis, and manual image annotation tools.

Machine learning. Machine learning is a subfield of artificial intelligence where statistical methods are used to “train” statistical models (e.g., logistic regression, support vector machines, or neural networks) from given data. In this context, data refer to any digital information, including sensor measurements, audio files, or graph structures. In this article we focus on data in the form of social media posts.

One common way to train machine-learning models on social media posts involves an assignment of individual posts to predefined categories or *labels*. Examples of labels include sentiment categories such as “negative,” “neutral,” and “positive,” or types of emotional reactions (e.g., “thumbs up,” “thumbs down,” etc.). In some

000 Journal of the Society for Social Work & Research Spring 2020

cases, labels are readily available on a social media website, but often, labels must be manually added to data—a task called annotation. In machine learning, a data set usually contains both data and labels. Training a statistical model on a social media data set requires updating the model's parameters according to specific rules that map social media posts to their associated labels as accurately as possible.

The ultimate aim is to obtain a model that is able to reliably predict labels. For example, Twitter posts (i.e., tweets) could be labeled by their sentiment as either negative, positive, or neutral. A model would then be fit on the labeled tweets with the goal of detecting the sentiment of new tweets. Another classic example would be to mark objects (e.g., a car) in images by a rectangle that contains it, called a *bounding box*. This task of manually adding bounding box information to images is referred to as bounding-box annotation. With a sufficient number of bounding-box annotations, a computational model can then be trained to predict bounding boxes in images for the given type of object.

Crowdsourcing. In the context of data analysis and annotation, crowdsourcing can be seen as an approach where items are annotated by *crowdworkers*—individuals who are registered on a crowdsourcing website such as Amazon Mechanical Turk to work on online annotation tasks for monetary incentives. In this approach, crowdsourcing websites serve as hubs between crowdworkers willing to work on paid annotation jobs and the people or institutions offering paid annotation jobs.

The growing importance of annotated data, in computer science in particular, presumably fostered the rise of many such websites, and crowdsourcing has become a well-established approach for annotating large amounts of data. This is especially appreciable in the context of machine learning, where several well-known data sets have been built mainly through crowdsourcing (e.g., ImageNet by Deng et al., 2009; Microsoft COCO by Lin et al., 2014; and Visual Genome by Krishna et al., 2017), and votes from crowdworkers have been used for various evaluation tasks such as assessing data set quality (Zhao, Yao, Gao, Ding, & Chua, 2016) or estimating the quality of model predictions (Mathews, Xie, & He, 2016).

There are some important differences between working with crowdworkers and in-house annotators. As the term suggests, crowdsourced annotators are generally seen as a “crowd”—an anonymous mass that is not known personally and might be distributed worldwide. Typically, there is no direct communication (e.g., direct messaging or calls) between the researchers conducting the project and the crowdworkers providing the annotations. Doing so might even be prohibited, as it is by Amazon Mechanical Turk. Consequently, the relationship between researcher and annotator is generally less direct in crowdsourced studies, and the only communication is via the annotation task. This implies that results are not discussed with the crowdworkers, and perhaps more importantly, this distance is likely to impact annotator motivation. Another characteristic of crowdsourcing has a clear effect

on motivation: Annotators are generally paid per annotation, which means that the crowd is inclined toward completing annotations quickly.

Although crowdsourcing might be a reasonable approach to quickly complete some noncritical tasks—such as labeling whether images show cats or dogs—one has to keep in mind that crowdworkers are used to tasks of short duration and tend to perform better in simple annotation scenarios. Even in such cases, however, crowdsourced annotations are not perfectly reliable. The literature commonly assumes that unreliable annotations stem from unethical spammers who submit imprecise or arbitrary labels in order to maximize their financial efficiency, malicious workers who purposefully aim to undermine or influence the labelling effort, and unqualified workers.

It is important to note that unqualified workers are, despite their best efforts, unable to produce an acceptable annotation quality (Eickhoff, 2018). The lack of expertise is more relevant for marginalized communities or other critical domains that require specific training or background knowledge that are common in the field of social work. Poor performance is subsequently propagated into machine-learning models, as the models statistically fit the resulting data set with the purpose of learning to label samples the same way it was done by annotators. As a result, unreliable annotations can lead to models with low classification accuracy and biased predictions. This issue is why social work should drive social media annotation and interpretation of data and results, particularly when it relates to the most challenging social problems.

Because such quality issues are well known, several mechanisms for increasing quality are common in crowdsourcing, such as collecting multiple annotations for each sample, or excluding annotators who complete the task in an unreasonably short time. Another option is that researchers specify which answers are acceptable for a small subset of the data and then exclude annotators whose proportion of acceptable answers is too low. Still, multiple opinions per sample do not necessarily reveal systematic interpretation biases (e.g., due to missing domain expertise), and crowdsourcing platforms are not designed for in-depth annotator training. Moreover, it can be hard to have the necessary degree of control over the ephemeral crowd workforce when relying on common websites such as Amazon Mechanical Turk (Difallah et al., 2014). Hence, we advocate selecting and training few annotators properly rather than relying on many untrained individuals, especially when analyzing social media or when annotation biases might have critical implications.

In principle, on some crowdsourcing platforms it is possible to work with selected and trained in-house annotators by creating a private link to access the task and only sharing this link with certain people. But such crowdsourcing platforms would typically still need to be paid, and, more importantly, do not provide the level of flexibility we desired. For example, it can be challenging to display social media posts in

000 Journal of the Society for Social Work & Research Spring 2020

the same or similar style as on the original social media platform, to distinguish between experts and nonexperts, and to enact precise control over how items are sampled for annotation.

Open-source projects for crowdsourcing. *Open-source* refers to computer software with publicly accessible source code that is released under a permissive license, allowing others to modify and share the software (Laurent, 2004). Typically, open-source projects make their code available via websites such as GitHub or GitLab, where others can download and contribute to the code. Because source code can be accessed and modified freely, open-source software generally offers a large degree of flexibility. However, adapting open-source software often requires more technical proficiency than commercial solutions, especially if code-level changes are necessary.

There are several open-source projects for crowdsourcing software, including *The New York Times* R&D Lab's "hive" (<https://github.com/nytlabs/hive>), ProPublica's "Transcribable" plugin (<https://github.com/propublica/transcribable>), Zooniverse's "Scribe" (<https://github.com/zooniverse-glacier/Scribe>), and Scifabric's "PYBOSSA" (<https://github.com/Scifabric/pybossa>). However, with the exception of PYBOSSA, these projects would require significant customization to process social media data, and their source code does not seem to be maintained.

Regarding the latter project, PYBOSSA is designed for building crowdsourcing websites where crowdworkers can register and work on available tasks. PYBOSSA is likely to be adaptable for having in-house annotators instead of crowdworkers do all annotations, but its source code consists of over 45,000 lines of code; we estimated that adapting such a complex framework to our case would likely amount to more work than building a new lightweight system. (For comparison, the complete VATAS source code has around 3,000 lines.) Several annotation websites do build on PYBOSSA (e.g., <https://crowdcrafting.org/>) and can be used to collect annotations; however, this brings about the same problems outlined in the previous section.

Qualitative data analysis. Qualitative methods are a core component of social work research. The tenets of qualitative methods take a more person-centered approach, privileging context, depth, a holistic perspective, and inductive rather than deductive reasoning. This is particularly important when coding social media data where misinterpretation of text or bias in labeling could lead to the criminalization of users, particularly individuals from marginalized backgrounds (Frey et al., 2018; Patton et al., 2017). There are several qualitative data management systems available that allow for social media use. For example, Dedoose is a fee-for-service, Web-based system that allows the researcher to import and analyze social media data. Nvivo provides a similar fee-for-service option, but it is only accessible on Windows and Mac operating systems. DiscoverText allows researchers to analyze unstructured social media, providing the user with control over the parameters of the data analysis to fit their research questions.

However, Dedoose and Nvivo were not designed for social media data, and Nvivo does not provide a user-friendly platform for naturalistic annotations. In addition, it is difficult to switch back and forth from the annotation task to the original webpage of the social media post. Nvivo requires that software be installed, which might create access challenges for annotators who require technological support. Although DiscoverText offers innovative features for Twitter analysis, the platform is less user friendly if coding or analyzing text with communities or organizations who may be less familiar with analysis software.

Manual image annotation tools. Many machine-learning approaches rely on annotated data to build computational models. For images in particular, there are various common types of annotation that are each relevant for building different types of machine-learning models. For example, during annotation, images can be manually assigned to predefined categories, or each pixel of an image can be assigned to a category (e.g., “house,” “car,” “street”).

In the current study, we were particularly interested in collecting bounding-box information (i.e., rectangles around objects of interest, in tandem with information about which class an object belongs to). See Figure 2 for an example of specific bounding-box annotations. Many tools exist for this annotation task (e.g., LabelMe by Russell et al., 2008; IAT by Ciocca, Napoletano, & Schettini, 2015; and Accurator by Dijkshoorn, Boer, Aroyo, & Schreiber, 2017), but there are two major problems with most of these tools. First, it is often impossible to display text with the image or to customize the user interface in other ways for displaying complete social media posts. Second, it can be hard to extend or integrate these tools into a more comprehensive annotation system where annotators log in and view the original tweet on Twitter, do bounding box annotation, and answer a list of other questions. One particular code repository for marking bounding boxes in images, the bbox-annotator of Kota Yamaguchi (code available at <https://github.com/kyamagu/bbox-annotator>), included most of the functionality we desired for marking concepts inside images. We used this code for the bounding-box component of our annotation system.

VATAS Annotation Tool

Annotation Tool Development Process

Our research team initially qualitatively analyzed tweets from gang-involved youth using Excel spreadsheets to capture text and emojis. This process was inefficient and made it difficult to visualize the data from a dynamic, naturalistic perspective. To counter this limitation, we developed a systematic approach to analyzing pathways to violence on Twitter among gang-involved youth that places the annotator in the shoes of a Twitter user. Having access to previous Twitter posts, the user’s social

000 Journal of the Society for Social Work & Research Spring 2020

network, images, and conversation provides important contextual clues about how content becomes aggressive on Twitter.

We formed an interdisciplinary team of social work researchers and computer scientists and created a list of visual concepts (e.g., guns or hand gestures that are visible in some images associated with the tweets) that would be useful for identifying pathways to violence on social media and could later be detected automatically. For this project, we wanted to select and train annotators to annotate tweets with respect to the visual concepts and a list of communication codes and answer additional qualitative questions about the items. These annotations were collected with two goals in mind. First, we wanted to use annotations to build automatic detection methods with limited misinterpretations involving biases (e.g., biases due to lack of contextual data). Second, we wanted to derive insights for social work research and practice.

We met weekly to discuss creating a Twitter data set consisting of text and images with the goal of improving detection of responses to loss and aggressive posts among youth in Chicago, IL, neighborhoods with high rates of community violence. The creation of a visual ontology of tweets related to loss and aggression would require that some tweets be manually annotated. We wanted a more robust annotation process that would go beyond labeling tweets as either aggression- or loss-related and would include the capability to annotate bounding boxes around concepts in associated images. The result was VATAS, a system suitable for private annotation (i.e., only having experts from the community and research assistants from our group annotate the data) that is capable of extracting deeper contextual meaning, culture, and complex nuance embedded in and around the tweet. Creating VATAS for annotation also had additional advantages, such as increased privacy and maximal flexibility regarding system functionalities.

Key Features of VATAS

VATAS is an open-source software for building websites for social media data annotation; it was designed specifically for cases that require a deeper understanding of contextual information, such as domain expertise. The complete code and technical details can be found on GitLab: <https://gitlab.com/blandfort/VATAS>. The key features of VATAS include the following:

- Intuitive annotation: This is particularly important when working with domain experts who might have little technical knowledge and limited time to get accustomed to the annotation process.
- Web based: Annotators do not have to install anything and can provide annotations from anywhere via a Web browser.

VATAS 000

- **Flexibility:** The system is open-source and fully customizable. In particular, the researcher has full customizability control over annotation tasks and ordering of items within each task, and they can implement more complex system behaviors (e.g., moving to annotation Task B if a certain response was given while working on Task A). Customizable layout templates are used to display tasks for annotators, and these templates can be shared across tasks.
- **Annotator roles:** For each task, any VATAS user can be assigned as a “normal” annotator or as a domain expert. For each of the two groups, the number of annotations to collect per item can be separately specified.
- **User management:** Each VATAS user can have either standard or administrative rights. This distinction is used to decide which tasks a user can work on and which annotations can be viewed and edited. All VATAS users can work on annotation tasks they have been assigned to and view and edit their completed annotations. Administrators can access all annotation tasks and view and edit annotations of all other users.
- **Handling annotation conflicts:** Administrators can view conflicting annotations and break ties by providing an extra annotation.
- **Export annotations:** Annotations for any task can be directly exported as a text file with tabular data (i.e., CSV). In this process, annotations are automatically paired with their corresponding social media data.
- **Privacy:** Users can host the system themselves, so no data are shared with any third party.
- **Free:** VATAS is free for commercial or noncommercial purposes.

VATAS Workflow

Detailed instructions for setting up VATAS and adding annotation tasks can be found on GitLab. Here, we outline the basic workflow for running an interdisciplinary study involving social work and computer science researchers.

Project goals. To establish the scope of the project, the team should decide on the primary research questions or goals. We recommend having one question pertaining to social work and one pertaining to computer science, and questions or goals should be related to and benefit from each other. For example, computer scientists might use insights from qualitative analysis to improve detection methods, whereas social workers might use detection models for practical applications or to find additional data more easily.

Data collection. Generally, the social work team should formulate criteria for collecting suitable social media data (e.g., identifying seed users for snowball sampling). The computer science team can then implement and run the data collection.

000 Journal of the Society for Social Work & Research Spring 2020

VATAS setup. The user needs access to a Web server where VATAS can be hosted. The computer science team sets up the annotation system on the server, including downloading the latest version of the VATAS source code, creating a database, and adjusting the settings. (Details can be found on the project's GitLab page.)

Annotation tasks. Designing annotation tasks requires the inclusion of people with technical knowledge (typically computer scientists) as well as those with domain knowledge (typically social work researchers) in discussions to ensure that it will be feasible to implement the final tasks with reasonable effort while keeping the process informed by domain expertise, and ensuring that the research interests of both groups can be satisfied by the final tasks. Adding the tasks to VATAS should be done by the computer science team, as it involves technical steps on the server side. (See the project's GitLab page for details.)

Training. The social work team trains annotators on VATAS use, ways to approach each task, and the ethics of annotating social media data in the VATAS system (e.g., uncovering their annotation biases, annotating in a private space, and confidentiality).

Annotation process. Annotators work on tasks, and the social work and computer science teams monitor incoming annotations and reconcile disagreements. We recommend iterative discussions between annotators and social work administrators throughout the annotation process.

Data export and analysis. In VATAS, any user with administrative permissions can download all annotations for individual tasks in CSV format. Exported annotations are analyzed by the social work team, and the computer science team uses the annotations for training detectors and/or running statistical analyses. Both teams should jointly discuss their findings.

Case Study: Chicago Twitter Corpus

To illustrate how VATAS can be adopted for rigorous study, we describe a collaborative project that used the tool for annotation and analysis of social media data from a marginalized community. Even though VATAS was developed during this project, to make it easier for the reader to transfer the process to his or her individual case, we are writing this manuscript from the perspective that the code for VATAS was already created. We received an institutional review board exception for this study because all of our social media data is publicly available.

Project Goals

In our research, we asked: How do Black and Latinx youth living in neighborhoods with high rates of community violence respond to loss and express aggression on social media? To answer this question, we assembled an interdisciplinary team that included social work researchers and students, computer scientists, youth, and outreach workers. Interdisciplinary questions included: "How does online aggression

lead to offline violence?” on the social work side and “How can we use machine learning to detect online aggression?” on the computer science side. Together, we annotated social media interactions among young people, which unearthed root causes of violence (e.g., poverty, trauma) and provided the training data for machine-learning analysis used to predict behavioral patterns on social media.

Data Collection

To obtain social media data from these populations, we started with a seed user on Twitter who self-identified as gang involved, had a large Twitter following, and whose story (including her death) was nationally covered in the media. We then identified her top communicators on Twitter through replies and mentions. Using her Twitter account and the accounts of her top communicators, we used snowball sampling techniques (Atkinson & Flint, 2001) to find other Twitter users who were from Chicago and expressed similar experiences of self-identified gang involvement or experiences navigating violence in order to build our social media corpus.

For our social media corpus, we identified 279 unique Twitter users living in Chicago neighborhoods with high rates of violence. We created two data sets: a text social media corpus and an image data set. For our text social media corpus, we collected the last 200 tweets for each unique user (or less depending on how many tweets each user had); for our image data set, we collected 1,851 tweets with images, randomly sampling 173 users from our total sample. There was no overlap between tweets in the social media text corpus and in the image data set, even though the text corpus does contain some tweets with images.

VATAS Setup

The computer science team installed VATAS on the university server as specified in the instructions on the project’s GitLab page. Our research project required various roles to keep VATAS and the annotation process running smoothly. The system administrator was consulted iteratively to make sure the servers hosting VATAS were running properly and efficiently. Leaders of the social work team (a social work professor and doctoral student) met weekly to discuss changes to the VATAS system based on new insights from annotations and communicated those needs to the computer science team. Computer scientists on our team were responsible for adapting and revising annotation tasks, roles, and permissions as they received feedback from the social work team.

Designing and Adding Annotation Tasks

Once the annotation system was set up, we then designed annotation tasks and added them to the system. After conversations between the social work and computer science teams on the research questions and types of analyses, annotation tasks were developed and organized to meet the needs of each research team. We designed four

000 Journal of the Society for Social Work & Research Spring 2020

annotation tasks: full-text annotation, collapsed-code text annotation, full-image annotation, and collapsed-code image annotation (see Table 1). We created full annotation tasks to capture detailed, robust annotations used for descriptive and thematic qualitative data analysis. Full annotation tasks (e.g., researching a word with a local community context and meaning) have many questions and require focus and time, whereas collapsed-code annotation tasks are for the training and development of computational systems that automatically detect textual labels and visual concepts. Collapsed annotation tasks are quick and require extensive domain and subject/content-specific knowledge of violence, social media language, and context (e.g., extensive knowledge of relationships between various sets of gangs and crews) in order to minimize annotation error.

For all annotation tasks, we started by supplying a social media post from our corpus. The text and/or image was available to each annotator as well as a link to the original post online. These posts may include hashtags, emojis, links, images, and videos. In the full-text and image annotation tasks, annotators analyzed all posts from each unique user in chronological order, whereas annotators for the collapsed annotation tasks were shown posts at random. The supplied social media post (and image) remained accessible on the left part of the screen for every question of each task. For all tasks, each social media post was annotated by at least two different annotators, and sometimes more.

The rest of this section describes each task, including the instructions we gave annotators. VATAS supports a wide variety of annotation functions, including high-level free-text interpretation based on domain knowledge; assignments of numerical ratings of codes generated by qualitative analysis; and object specifications, such as image bounding boxes. This diverse set of annotation functions can be easily customized to support other social media studies.

First impression (text and image). Both text and image annotation tasks began by asking annotators about their first impressions of a social media post (see Figure 1). If the post came up on their own social media feed, what would their initial interpretations be? What would come to their mind right after seeing it? We started with this to capture annotators' baseline interpretation, evaluate assumptions, and uncover biases that may affect how annotators see the post. Once our annotators had acknowledged these initial interpretations, assumptions, and biases, they could consider these throughout the rest of the tasks. This annotation step is especially important when analyzing data from marginalized and vulnerable communities where interpretations and labeling could lead to further harmful implications for the people in these communities.

Contextual analysis of social media instructions (text and image). Our discourse and textual analysis sought to uncover contextual information about each social media text and image, which included analyzing the various components of each. For this purpose, we developed the *contextual analysis of social media* (CASM) approach for

Table 1
Overview of Questions for the Four Annotation Tasks

Question	Input Format	Full-Text		Full-Image		Collapsed-Code	
		Annotation	Annotation	Text Annotation	Image Annotation	Text Annotation	Image Annotation
First impression	Open-ended	Yes	Yes	–	–	–	–
CASM instructions	–	Yes	Yes	–	–	–	–
Location	Multiple choice, short answer, and open-ended	–	Yes	–	–	–	–
General description	Open-ended	Yes	Yes	–	–	–	–
Threat level	Scale (0–1; intervals of 0.1) and open-ended	Yes	Yes	–	–	–	–
Lean* indicators	Scale (0–1; intervals of 0.1) and open-ended	–	Yes	–	–	–	–
Bounding boxes	Image concept bounding and labeling	–	Yes	–	–	–	–
Code	Multiple choice (multiple answers accepted)	Yes	Yes	–	–	–	–
Collapsed code (experts only)	Multiple choice (multiple answers accepted)	–	–	Yes	Yes	–	Yes

Note. CASM = contextual analysis of social media. *Lean is a drink that mixes promethazine with codeine and soda or juice.


Main Annotation Task
Annotation overview
Select task ▾
Logout (expert)

Tweet

@user2

Jay Smokin thinkin bout DMoney 🙏🙏🙏

2018-23-07 14:54



(View image in full size)

Part 1 – First Impression

If this photo came up on your social media wall, what would be your first impression? (What would come to your mind right after seeing it? How would it make you feel?)

This user is about to roll and blunt and smoke some weed, while they think about someone close to them who has passed away named "D Money"

Figure 1. Screenshot of full-image annotation task in VATAS (Visual and Textual Analysis of Social Media).

unpacking context in social media posts (Patton, Frey, McGregor, Lee, & McKeown, in press). CASM includes a deep dive into the original social media post, the user of the post, the peer network of the user, any offline events referenced, virality (likes and reposts), and engagement (comments and replies). We outlined each of these instructional steps for our annotators at the beginning of each group of tasks so these steps would not be forgotten. We asked our annotators to use websites, search engines, and various other resources to find the potential meaning of the social media post. Although the annotator did not need to directly input anything for this task, they would be unable to effectively finish the following tasks if this step was not thoroughly completed.

Location (image only). Within our group of image annotation tasks, we asked annotators to reflect on the location represented in the image. Where does the main event take place? Or if there is no event taking place in the image, what is the location of the main subject(s) or object(s) that are visible? We broke images into three locational categories: inside, outside, or other (e.g., images of text and memes). Once our annotators choose one of these categories, we asked them to write the precise location and to describe it. For example, if an annotator categorized an image as indoor, they might write “bathroom” and describe the features that led them to understand the image as being taken in a bathroom. We wanted to capture the location of images in order to analyze patterns regarding what is happening in images in various locations and any themes in the corresponding text (e.g., substance use and expressions of grief).

General description (text and image). Once annotators completed CASM, they were asked to reflect on and evaluate their initial interpretation of the post. This evaluation was completed through a synthesis of every contextual detail they found to describe the meaning of the social media post. They were tasked with providing the evidence that led to their interpretation of the post, including a Standard English rephrasing that fully captured the meaning of the social media post. For images, we asked annotators to describe what was happening in the image and how the image related to the text in the post (if there was any relation).

Threat level (text and image). After annotators described their baseline evaluation and final interpretation of the social media post, we asked them to comment on the threat level of the post. The annotators indicated threat level on a scale from 0 (*not at all threatening*) to 1 (*extremely threatening*) by increments of 0.1; they had the opportunity to provide an open-ended answer to explain the threat level they gave. We wanted annotators to specifically think about threat in terms of how likely a post would lead to someone being harmed or hurt. Although a post might display aggression, it could be ambiguous and not immediately threatening. Inversely, a post displaying aggression could be credible and specific, leading to a high threat level. We were interested in threat level to determine patterns in posts that lead an annotator to deem a post as threatening.

Lean indicators (image only). We were particularly interested in substance use in our data set, as altered states of mind while using social media could result in themes and patterns of certain content. *Lean*—a drink mixing promethazine with codeine and soda or juice—is used by young people in neighborhoods where our social media data originated. Therefore, we wanted annotators to note when lean showed up in images through containers of promethazine or through containers and cups with purple, red, or yellow liquid with clues in the social media post. Because we may not be completely sure that lean is present in an image, we gave our annotators a scale from 0 (*not at all likely*) to 1 (*completely likely*) by increments of 0.1. If the annotators ranked the image above a zero, they were required to provide their evidence for doing so in an open-ended answer box.

Bounding boxes (image only). We were interested in automatically detecting visual concepts—person, tattoo, hand gesture, firearm, money, marijuana (raw), joint/blunt/cigarette, and indications of lean—in images that were related to the users in our data set. To train VATAS to do this, our annotators had to manually annotate images by drawing bounding boxes around the visual concepts (see Figure 2). Once we had enough hand-annotated images, we were able to begin training VATAS to automatically detect these concepts in images that had not been manually annotated. We were also able to do comparative analyses on text content related to the concepts depicted in the images.

Code (text and image). Annotators were instructed to pick a qualitative code that best represented the essence of each social media post. They were able to select

The screenshot displays the VATAS interface. At the top, there are navigation links: "Main Annotation Task", "Annotation overview", "Select task", and "Logout (user)". The main content is divided into two columns: "Tweet" and "Context".

Tweet: The tweet text is "IF U Aint In da Streets Or Kno What Goin On Dont Speak On Whats Goin On" with a blue speech bubble icon. The user is "@user3" and the timestamp is "2018-23-07 19:20". Below the text is a photograph of a person's hands in a white t-shirt, with a green bounding box drawn around them. The text "HAND GESTURE" is written in green above the box.

Context: The text reads: "Before you answer the questions below, make sure you understand the tweet and are aware of its context:" followed by a numbered list:

1. Look at the text (emojis, hashtags)
2. Look at the image
3. Look at any other links in the tweet
4. Go to the original post on Twitter
 1. See who the poster is
 2. If anyone is @ed, see who they are
 3. See if any offline events are referenced
 4. See the amount of likes and retweets
 5. See if there are any replies/comments on the tweet

Part 4 – People and Objects

Please draw bounding boxes directly on the shown image to mark all of the concepts below that are displayed in bold. (Make sure you are familiar with the [general guidelines].)

- **person**
- **money**
- **firearm** [?]
- **lean container**
- **joint/blunt/cigarette**
- **marijuana** [?]
- **hand gesture**
- **tattoo**

Click the question marks to show additional information for the corresponding concepts.

Figure 2. Screenshot of bounding-boxes question in VATAS (Visual and Textual Analysis of Social Media).

more than one if necessary but were instructed to do their best to choose only one. Codes ranged from “growth,” “grieving,” and “aggression” to “health,” “mood,” and “social behavior.” We used the qualitative codes to track themes and patterns in the data set.

Collapsed code (collapsed text and image). Collapsed coding for text and image annotation is primarily used to speed up the process of labeling data for the training of computational detection methods. This task is reserved for subject/content expert annotators or domain expert annotators because of the foundational and domain-specific knowledge required to complete the task quickly with as few errors as possible. In our social media data set from young people in Chicago neighborhoods with high rates of violence, we were specifically interested in coding for aggression, loss, and substance use (see Figure 3; Blandfort et al., 2019; Blevins et al., 2016; Patton, McKeown, Rambow, & Macbeth, 2016).

Training

To start the annotation process, our annotators visited the annotation website on a personal laptop or tablet, logged in with their username and password, and selected the specific annotation task on which to work. Our study had two groups of annotators: Master of Social Work student research assistants (RAs) and expert annotators (subject/content experts and domain experts). RAs spent most of their

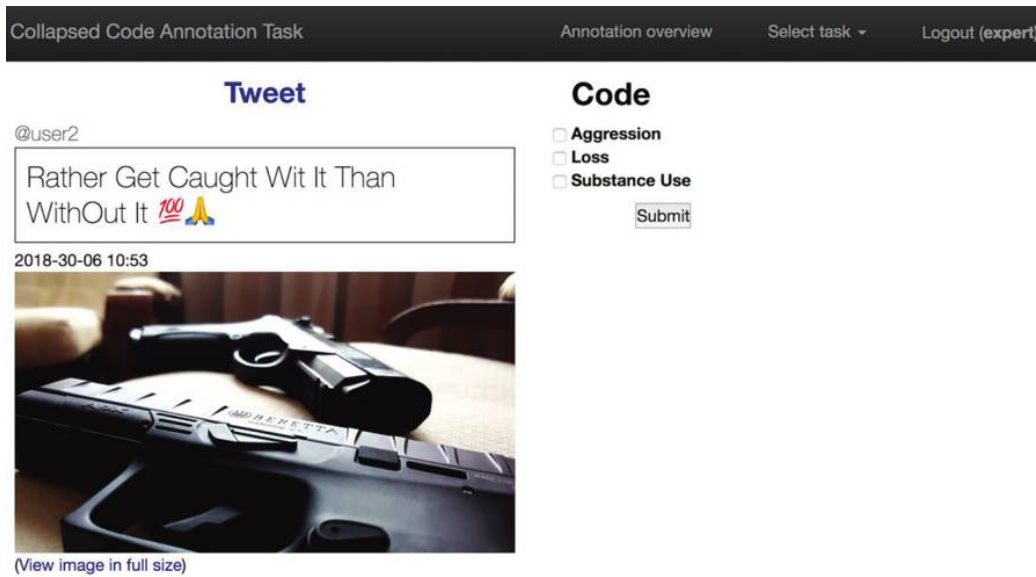


Figure 3. Screenshot of collapsed-code task in VATAS (Visual and Textual Analysis of Social Media).

time completing full-text and full-image annotation tasks, and expert annotators completed collapsed-code annotation tasks (text and image).

Although RAs brought their own expertise to their annotations—including their knowledge of violence prevention, youth development, social systems, and ecological frameworks—they needed to be trained in other areas before they could begin full annotations. The training included an overview of the domain where the social media data originated (Chicago, various crews/gangs, and geographical space), an explanation of their role as annotators of social media data, an extensive tutorial of VATAS, and process meetings to prevent bias toward certain groups or content during the annotation (Patton et al., in press).

Following this initial training, RAs engaged in a deep social media immersion by observing social media posts and learning the different functionalities of Twitter and the ways youth were communicating. Finally, RAs practiced annotating 100 social media posts to prepare them for the official data set (Patton et al., in press). Throughout the training, the trainer monitored each RA's progress and provided iterative feedback to improve annotation quality. After RAs completed the 2–3-week training, they were ready to begin full-text and image annotation tasks. RAs were tasked with completing 100 annotations per week; they met weekly with the trainer to discuss the annotation process and content, work through challenges, and discuss areas for improvement.

Expert annotators came in with extensive knowledge of the domain and content. Our content/subject expert annotators were a professor of social work specializing in social media and violence, and a social work doctoral student with more than 9 years of organizing, mentoring, and advising experience with youth of color and

000 Journal of the Society for Social Work & Research Spring 2020

3 years of experience with social media and violence. They completed both collapsed-code text and image annotations due to their extensive knowledge of violence, social media language, and contextual analysis. Content/subject expert annotators were also tasked with reconciling social media posts labeled by RAs. In VATAS, there was an indicator when RAs disagreed on the code given to a social media post. A subject/content expert could then go through these posts and reconcile the disagreement by choosing which label best matched their interpretation of the post.

We also hired domain experts to annotate posts; these were people residing or working in neighborhoods with high rates of violence and who had professional or personal experience with violence. Domain experts often had limited time to annotate for reasons including (but not limited to) school, familial obligations, community service, and work. Although they could complete all the tasks involved in our annotation tool, we wanted to make sure we were not demanding more of their time than they could offer. Domain experts in our specific study spent their time completing the collapsed-image annotation, which is a quick task where they were able to harness all of their domain expertise and provide labels for social media posts with images used to train VATAS.

Data Export and Analysis

Whenever the social work or computer science teams wanted to export completed annotations, an administrator went to the annotation website, logged into their account, and clicked on the export tab at the top of the page to download data as a CSV file. Administrators could export various groups of annotations depending on their needs: full-text annotations, full-image annotations, collapsed-code text annotations, or collapsed-code image annotations.

Once our team exported the CSV files containing the data annotated through VATAS, these files were used for qualitative analysis as well as machine-learning training and experiments. The social work team accessed CSV files through Excel for qualitative thematic analysis, looking for patterns in the annotated data (e.g., the frequency of social media posts referencing death preceding posts about having trouble sleeping) to inform violence prevention and intervention practices and adapt future annotation foci. The computer science team analyzed the CSV files in Python and used parts of the data for training and analyzing computational methods. In particular, bounding-box image annotations (together with the corresponding social media images) were used to train a computer-vision model to recognize the annotated visual concepts and collapsed-code annotations (together with the original tweets). This served as training and test data for classifying tweets as belonging to the categories *loss*, *aggression*, *substance use*, or *other*. These classifiers were then analyzed to find out which concepts in the texts and in the images were most indicative for the respective codes.

Discussion

Limitations

VATAS has several limitations. First, it was developed by researchers and is a non-commercial alternative to crowdsourcing platforms. Thus, the system is free to use but comes without technical support. Research teams using VATAS are responsible for its setup, configuration, and maintenance. As noted earlier, this requires certain technical skills, including at least a basic understanding of programming. Furthermore, a Web server is required to host VATAS. However, the technical skills related to developing and maintaining VATAS offer an opportunity for social work researchers to create new research collaborations and partnerships with computer science colleagues.

Second, we explained earlier why crowdsourcing would not be a viable option for our investigation. However, in some cases when the annotation task is simple enough, crowdsourcing might be more appropriate because annotations can be collected more quickly due to large numbers of available crowdworkers. In particular, our general annotation approach of having manually selected and trained annotators perform rigorous analysis for all items might not be directly applicable to the annotation of large-scale data sets.

Ethics

VATAS requires not only technical and methodological considerations, but ethical ones as well. Ethical considerations include the ways social media research could directly or indirectly impact study populations, clarifying the ethical obligations specific to each population to ensure that the research does not cause further marginalization or harm, and adopting mechanisms to protect the study population (e.g., privacy). Although we only used public tweets, our work has the potential to draw more attention to users—in our case, Black youth who may already face marginalization, criminalization, and surveillance, both online and offline.

Social work researchers are offered little ethical guidance when seeking to engage novel methods for conducting research using social media data. The National Association of Social Workers (2017) Code of Ethics does not provide guidance on the ethics of social media research. In 2017, the National Association of Social Workers, the Association of Social Work Boards, the Council on Social Work Education, and the Clinical Social Work Association released a report on technology in social work practice, which only briefly mentions social media and conducting online research. Furthermore, institutional review boards often lack clear guidelines on obtaining consent from social media users with public accounts; in some instances, coding social media data is exempt from full review.

000 Journal of the Society for Social Work & Research Spring 2020

Despite a lack of institutional and organizational guidance regarding ethics, when using VATAS we strongly recommend a rigorous ethical review process and consulting institutional review boards, human research protection specialists, and leaders from the community where social media data originates to ensure that the safety and protection of social media users remains central in the research. To protect users from potential harm caused by our research and use of VATAS, we de-identified each social media post in this article, thus rendering the text unsearchable. De-identifying social media posts involves removing identifiable information in each post (e.g., social media username, replacing the images, and replacing names with pseudonyms). Due to Twitter Advanced Search, a feature through which any tweet can be searched by user, text, and date, we rendered each tweet unsearchable by altering the text of the tweet while not jeopardizing meaning. We then tested whether the tweet was truly unsearchable by iteratively searching the newly formed tweet piece by piece and in its entirety to see if the original is found. All of the images presented in this manuscript are from Flickr: Creative Commons and not from our data set. We share our data set and computational tools only with partner organizations and other researchers who sign a memorandum of understanding outlining their research purpose(s) and intention(s). Using VATAS for social media data annotation and analysis requires a deep understanding of our ethical obligations for preventing potential risks to our involuntary participants.

Conclusion

As innovative new ways of communicating, sharing, and connecting continue to diffuse throughout society, our methods for understanding and making meaning of data must advance as well. Although these modalities rely heavily on technology, at their core they remain uniquely human in their intricacies with dependent and fluctuating contextual factors, as well as unique cultural components. Taking cues from the data science world, social work researchers need to develop their own innovative methods to capture robust social media data that can work synergistically with data science efforts. There are technical hurdles to overcome that social work research teams may be unfamiliar with, further underscoring the necessity of highly collaborative and mutually beneficial relationships with data science teams.

Although technology and data science will continue to advance our ability to use and make sense of text, images, audio, and video data, there is a fundamental need for interdisciplinary collaboration to make sense of this information. Qualitatively informed features of linguistically unique text provide insight helpful to development of automated systems and provide qualitative researchers with opportunities to develop in-depth understanding of social phenomena in new naturalistic settings. At its best, this process is a mutually beneficial, multidisciplinary process where in-depth insights driven by social work researchers create new “soft” features to be used as a part of learning/training models to not only improve the accuracy of data

science models but also improve the accuracy, efficacy, and ethical use of machine learning and artificial intelligence. VATAS is one of the first steps in developing this new methodological toolbox required to use and make meaning of the new data streams available to social work researchers.

Critical to the development of VATAS was the interdisciplinary collaboration among social work researchers, domain experts, and computer scientists. We argue that this unique collaboration yields better science, producing a deeper understanding of the social phenomenon and an ability to more precisely measure a social problem and thus achieve greater impact. Social work research, as a field, has the opportunity to use its deep understanding of context, culture, and relationships to communities to inform new methodological and technological models that may lead to social change. Social media data provide social work researchers with new opportunities to learn more about their target populations in a naturalistic environment. Tools such as VATAS support rigorous and meaningful advancements in social work and social media research by incorporating mechanisms to capture social and cultural context when interpreting social media. Enhanced collaborations from disparate fields may unlock the next generation of great social work, epidemiological, sociological, and psychological research. VATAS facilitates communication among multidisciplinary teams and helps build truly meaningful partnerships among fields with dissimilar professional languages and approaches to explain social phenomena that may otherwise go unnoticed.

Author Notes

Desmond U. Patton, PhD, is an associate professor in the Columbia University School of Social Work and Department of Sociology.

Philipp Blandfort, MSc, is a doctoral student in the TU Kaiserslautern Department of Cognitive Science and Psychology and at the German Research Center for Artificial Intelligence.

William R. Frey, MSW, is a doctoral student in the Columbia University School of Social Work.

Rossano Schifanella, PhD, is an assistant professor in the University of Turin Department of Computer Science.

Kyle McGregor, PhD, is an assistant professor in the Department of Child and Adolescent Psychiatry at NYU Langone Health.

Shih-Fu U. Chang, PhD, is the Richard Dicker Professor in the Department of Computer Science and Electrical Engineering in the Columbia University Fu Foundation School of Engineering.

Correspondence regarding this article should be directed to Desmond Upton Patton, 1255 Amsterdam, New York, NY 10027, or via e-mail to dp2787@columbia.edu

Acknowledgments

Philipp Blandfort is co-first author of this manuscript. During some of this work, Blandfort was at Columbia University on a research stay supported by a fellowship within the German Academic Exchange Service FITweltweit program. Blandfort also received financial support from the Center for Cognitive Science, Kaiserslautern, Germany.

References

- American Academy of Social Work and Social Welfare. (2018). *Grand challenges for social work: Harness technology for social good*. Retrieved from <http://aaswsw.org/wp-content/uploads/2015/12/180604-GC-technology.pdf>
- Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social Research Update*, 33(1), 1–4. Retrieved from <http://sru.soc.surrey.ac.uk/SRU33.html>
- Blandfort, P., Patton, D. U., Frey, W. R., Karaman, S., Bhargava, S., Lee, F. T., . . . McKeown, K. (2019, July). Multimodal social media analysis for gang violence prevention. *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, No. 1, pp. 114–124). Retrieved from <https://www.aaai.org/ojs/index.php/ICWSM/article/view/3214>
- Blevins, T., Kwiatkowski, R., Macbeth, J., McKeown, K., Patton, D., & Rambow, O. (2016). Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2196–2206). Retrieved from <https://www.aclweb.org/anthology/C16-1207>
- Ciocca, G., Napoletano, P., & Schettini, R. (2015, February 18). IAT-image annotation tool: Manual. *arXiv:1502.05212* [cs.CV]. Retrieved from <https://arxiv.org/abs/1502.05212>
- Coulton, C. J., Goerge, R., Putnam-Hornstein, E., & de Haan, B. (2015). *Harnessing big data for social good: A grand challenge for social work* (Grand Challenges for Social Work Initiative Working Paper No. 11). Cleveland, OH: American Academy of Social Work and Social Welfare. Retrieved from <https://grandchallengesforsocialwork.org/wp-content/uploads/2015/12/WP11-with-cover.pdf>
- Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). doi:10.1109/CVPR.2009.5206848
- Difallah, D. E., Catasta, M., Demartini, G., & Cudré-Mauroux, P. (2014, September). Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*. Retrieved from <https://www.aaai.org/ocs/index.php/HCOMP/HCOMP14/paper/view/8958>
- Dijkshoorn, C., De Boer, V., Aroyo, L., & Schreiber, G. (2017, September 26). Accurator: Niche-sourcing for cultural heritage. *arXiv:1709.09249* [cs.CY]. Retrieved from <https://arxiv.org/abs/1709.09249>
- Eickhoff, C. (2018). Cognitive biases in crowdsourcing. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 162–170). doi:10.1145/3159652.3159654
- Frey, W. R., Patton, D. U., Gaskell, M. B., & McGregor, K. A. (2018). Artificial intelligence and inclusion: Formerly gang-involved youth as domain experts for analyzing unstructured Twitter data. *Social Science Computer Review*. <https://doi.org/10.1177/0894439318788314>
- Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3), 241–251. <https://doi.org/10.1016/j.bushor.2011.01.005>
- Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures & their consequences*. Thousand Oaks, CA: SAGE Publications.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., . . . Bernstein, M. S. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1), 32–73. <https://doi.org/10.1007/s11263-016-0981-7>

- Laurent, A. M. S. (2004). *Understanding open source and free software licensing: Guide to navigating licensing issues in existing & new software*. Sebastopol, CA: O'Reilly Media.
- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014, September). Microsoft COCO: Common objects in context. *European Conference on Computer Vision* (pp. 740–755). https://doi.org/10.1007/978-3-319-10602-1_48
- Mathews, A. P., Xie, L., & He, X. (2016). SentiCap: Generating image descriptions with sentiments. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 3574–3580). Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPaper/12501>
- National Association of Social Workers. (2017). *Code of ethics*. Retrieved from <https://www.socialworkers.org/About/Ethics/Code-of-Ethics/Code-of-Ethics-English>
- Patton, D. U., Brunton, D. W., Dixon, A., Miller, R. J., Leonard, P., & Hackman, R. (2017). Stop and frisk online: Theorizing everyday racism in digital policing in the use of social media for identification of criminal conduct and associations. *Social Media + Society*, 3(3), 2056305117733344. <https://doi.org/10.1177/2056305117733344>
- Patton, D.U., Frey, W., McGregor, K., Lee, F., & McKeown, K. (in press). Contextual analysis of social media: A qualitative approach to eliciting context in social media posts with natural language processing. *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*.
- Patton, D. U., McKeown, K., Rambow, O., & Macbeth, J. (2016, September 28). Using natural language processing and qualitative analysis to intervene in gang violence: A collaboration between social work researchers and data scientists. *arXiv:1609.08779* [cs.CY]. Retrieved from <https://arxiv.org/abs/1609.08779>
- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and Web-based tool for image annotation. *International Journal of Computer Vision*, 77(1–3), 157–173. <https://doi.org/10.1007/s11263-007-0090-8>
- Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, P., . . . Li, L. (2016). YFCC100M: the new data in multimedia research. *Communications of the ACM*, 59(2), 64–73. doi:10.1145/2812802
- Zhao, S., Yao, H., Gao, Y., Ding, G., & Chua, T. S. (2016). Predicting personalized image emotion perceptions in social networks. *IEEE Transactions on Affective Computing*, 9(4), 526–540. doi:10.1109/TAFFC.2016.2628787

Manuscript submitted: August 12, 2018

First revision submitted: October 27, 2018

Second revision submitted: November 2, 2018

Accepted: December 12, 2018

Electronically published: February 6, 2020

Supplemental Material (not copyedited or formatted) for: Desmond U. Patton, Philipp Blandfort, William R. Frey, Rossano Schifanella, Kyle McGregor, Shih-Fu U. Chang. 2020. "VATAS: An Open-Source Web Platform for Visual and Textual Analysis of Social Media." *Journal of the Society for Social Work and Research* 11(1). DOI: <https://doi.org/10.1086/707667>.

Appendix

Glossary of Terms

Terms	Definition
Amazon Mechanical Turk, CrowdFlower, FigureEight	On crowdsourcing websites such as Amazon Mechanical Turk, CrowdFlower and FigureEight, annotation tasks are set up for crowdsourcing and annotations from crowdworkers are collected.
bounding boxes	Bounding boxes are a type of annotation performed by drawing a box around specific objects in an image and assigning a label to the object (often used to build training sets for computer vision).
code repository	A code repository is a collection of code and text files associated with a programming project.
crowdsourcing	Crowdsourcing is a data annotation approach that uses crowdsourcing websites such as Amazon Mechanical Turk, where registered crowdworkers are paid for providing annotations.
crowdworker	A crowdworker is an individual who works on crowdsourcing tasks via a Web browser for monetary incentives.
domain expertise	Domain expertise is knowledge or experience in a specific field, which can refer to a particular community (e.g., Chicago gangs), social media website (e.g., Facebook, Twitter), or application area (e.g., health care, education).
emoji	Emoji are special symbols widely used inside text messages on social media platforms, often to express emotions. They include smileys such as 😊 😂 😭, as well as symbols such as ❤️ 🍕 🎉.
GitHub, GitLab	These are websites where code repositories can be hosted. They are widely used for sharing and maintaining code for open-source projects.
hashtag	A hashtag is a word-like structure that begins with the symbol “#” directly followed by a theme; hashtags are frequently included in social media posts (especially on Twitter) to signal that the post relates to the theme indicated by the hashtag. This can provide contextual information for interpreting the post and also makes it possible to find posts relating to the same theme. For example, in the post “goodbye future! #politics,” the hashtag “#politics” indicates that the author’s pessimistic view expressed by “goodbye future!” is related to politics.
metadata	Metadata is any information about data. For example, if the data consists of social media posts, metadata could include information about authors of the individual posts (e.g., user IDs, names, relations to other authors), information about how the data was collected (e.g., when the data was collected, which social media website the data was taken from, etc.), or annotations.

Supplemental Material (not copyedited or formatted) for: Desmond U. Patton, Philipp Blandfort, William R. Frey, Rossano Schifanella, Kyle McGregor, Shih-Fu U. Chang. 2020. "VATAS: An Open-Source Web Platform for Visual and Textual Analysis of Social Media." *Journal of the Society for Social Work and Research* 11(1). DOI: <https://doi.org/10.1086/707667>.

open-source	Open-source refers to computer software with publicly accessible source code that is released under a permissive license, allowing others to modify and share the software with anyone else.
server	In this article, we use the term server to refer to a computer that hosts websites (i.e., stores website content and sends it to Web browsers such as Google Chrome).
social media data set	This is a collection of social media data, with or without additional metadata (e.g., source, time, or annotations).
source code	Source code is a human-readable description of a computer program that consists of instructions and statements and that is written in a programming language.
spammer	This is a person who deliberately submits useless information (“spam”), which can include posting misleading, incorrect, or irrelevant content or spreading unsolicited advertisements.

Annotating Twitter Data from Vulnerable Populations: Evaluating Disagreement Between Domain Experts and Graduate Student Annotators

Desmond U. Patton
Columbia University
dp2787@columbia.edu

Philipp Blandfort
DFKI
philipp.blandfort@dfki.de

William R. Frey
Columbia University
wf2220@columbia.edu

Michael B. Gaskell
Columbia University
mbg2174@columbia.edu

Svebor Karaman
Columbia University
svebor.karaman@columbia.edu

Abstract

Researchers in computer science have spent considerable time developing methods to increase the accuracy and richness of annotations. However, there is a dearth in research that examines the positionality of the annotator, how they are trained and what we can learn from disagreements between different groups of annotators. In this study, we use qualitative analysis, statistical and computational methods to compare annotations between Chicago-based domain experts and graduate students who annotated a total of 1,851 tweets with images that are a part of a larger corpora associated with the Chicago Gang Intervention Study, which aims to develop a computational system that detects aggression and loss among gang-involved youth in Chicago. We found evidence to support the study of disagreement between annotators and underscore the need for domain expertise when reviewing Twitter data from vulnerable populations. Implications for annotation and content moderation are discussed.

1. Introduction

Annotation is the process of providing metadata (e.g. deeper meaning, context, nuance) through the act of labeling language or other contents such as images or videos. Machine learning and natural language research has long relied on the robust annotation of social media data to examine and predict myriad human phenomenon [10, 12, 14]. In the context of machine learning, the annotation process typically involves assigning categories to items, which are then used to build computational models for detecting these categories [1, 9]. With an understanding that language is highly subjective, researchers in computer science have spent

considerable time developing new methods to increase the richness of annotation [10] and combine annotations stemming from multiple annotators [18, 21, 25] based on estimated reliabilities [14, 19]. Most of these efforts have focused on inter-annotator reliability, improving accuracy across annotators and reducing disagreement regarding how to interpret data [10], often without analyzing causes of disagreement [14, 18, 19, 21]. Furthermore, these methods assume that for each given item there is one “correct” label. However, when human annotators disagree when choosing a different label for the same post, one must consider if there actually is a single correct answer. In addition, if an annotator holds more contextual knowledge than another, should some patterns of disagreements be weighed more heavily than others [19]? To extend this idea, we build on the work of Brock [6] and Roberts [20] who underscore the importance of centering the perspectives, viewpoints, and epistemologies of vulnerable and marginalized communities when analyzing social media data.

On the other hand, there is a gap in research which examines the positionality who annotates the data (e.g. demographics, expertise, experience), how they are trained and the extent to which those characteristics impact how data is labeled and interpreted. A deeper focus on annotation is particularly important when analyzing data from vulnerable and marginalized populations on social media. Symbolic interactionism theory suggests that the ways in which we derive meaning is in response to an interpretive process based in our social interaction with others [5]. That is to say, the meaning of social media posts from African American youth in Chicago and how they should be interpreted is rooted in a nuanced understanding of the everyday activities, experiences, language and shared culture. As such, the expertise and training of the annotators are important when observing local concepts, gestures, innuendo, and other psycho-social scripts and

behaviors embedded in text and images on social media. For example, in her book “It’s Complicated”, danah boyd describes a young African American male high school student who loses his spot at Yale University because of images on his Facebook profile that were interpreted as being connected to gang involvement. Misinterpreting nuances in language, culture, and context can have detrimental consequences that lead to criminalization and further stigmatization of marginalized groups [7, 16]. Determining when and if something is inappropriate is highly subjective and at the whim of annotators and content moderators who may have no familiarity with the language, concepts, and culture of the social media user [20].

In this paper, we present findings from the analysis of annotation within and between two groups: two African-American Chicago-based domain experts and two social work graduate students (one African American, one White) who annotated a total of 1,851 tweets with images from Twitter that are a part of a larger corpora associated with the Chicago Gang Intervention Study, which contains tweets with images from African American youth and young adults (See section 4). The broader purpose of this study is to develop a computational system that detects pathways to violence among gang-involved youth in Chicago. The paper is organized as follows. Section 2 provides a description of the annotation process. Section 3 provides a description of the methods for analysis of annotator perspectives. Section 4 introduces the case study which includes an analysis of differences in annotation within and between groups, what is revealed from those differences, and what to take from it. Section 5 describes implications from the study which include the importance of annotator training, how annotation should be monitored to identify problems, what to do with errors in annotations and how domain experts should be involved in the annotation process. Section 6 describes future directions which include other applications of our analysis methods and the implications of this work for content moderation.

2. Description of Annotation Process: The Contextual Analysis of Social Media Approach

The annotation process involves labelling tweets with respect to the psychosocial codes *aggression*, *loss*, and *substance use*, and contains various key components: annotators (Chicago-based domain experts and social work graduate students), social work graduate student annotator training, the Contextual Analysis of Social Media (CASM) approach [17], and a web-based visual and textual analysis system for

annotation [15]. The annotation process for each group of annotators has distinctions due to their different expertise.

2.1. Chicago-based domain experts

In order to ensure an accurate and contextual understanding of the images and text embedded in our Twitter corpus, we partnered with a local violence prevention organization in Chicago to hire two individuals as domain experts. We asked the partner organization to identify individuals who had a deep understanding of the local language, concepts, gang activity, and who were active on Twitter. The partner organization identified one African American man in his early 20’s, a community member, and one African American woman in her late 20’s, an outreach worker for the organization. The domain experts were asked to annotate 1,851 images using the annotation system. A white postdoctoral research scientist, with a doctorate in clinical psychology and based in Chicago trained the domain experts how to use the system, validated their community expertise, and clarified the purpose of the tasks and research. The domain experts were not trained on how to define and interpret *aggression*, *loss*, and *substance use* because we intentionally center their knowledge of community, language, and experience as expertise. As such, the domain experts are educating the researchers on how to define the aforementioned classifications [8]. Domain experts annotated the entire dataset on average within 48 hours from receiving the data because of their facility with the language and content embedded in the Twitter posts.

2.2. Social work graduate students

Social work graduate student annotators were current students in a Master of Social Work program. Both students are women and in their early 20’s one is African American and the other is White. They were chosen based on their professional experience in adolescent development, criminal justice, and community work with youth of color. All annotators showed and expressed an openness and willingness to learn through their prior work and participation in the SAFElab. The annotators undergo a rigorous training process involving five steps: 1) a general overview of the domain, 2) the annotator role, 3) annotation process and system tutorial, 4) deep Twitter immersion, and 5) annotation practice and feedback. The social work annotators received this specific training because they lacked the life experience that would provide them a firm understanding of the local context and language,

which could potentially lead to gross misinterpretations of the Twitter posts [16].

The training begins with an overview of the domain informed by insights from domain experts, which includes geography, relevant historical information (e.g., relationships and rivalries between gangs), and data around violence and victimhood. After the students received an overview of the domain, we outline their role as an annotator of Twitter posts. This involves describing the purpose and aims of the work and an introduction to thematic qualitative analysis of text and images. Additionally, our annotators engage with the ethical and sociopolitical aspects they will come across during annotation (e.g., privacy and protection, Twitter data from marginalized communities, implications regarding race), which includes understanding their own relation to the Twitter data and the domain [17].

Next, students are taken through CASM in our web-based annotation system, which includes instructions on accurate and efficient use of the system. CASM is a team-based contextually driven process used to qualitatively analyze and label Twitter posts for the training of computational systems. CASM involves a baseline interpretation of a Twitter post, followed by a deep analysis of various contextual features of the post, the post's author, their peer network, and community context. A thematic label is then applied to the post. These reconciled labeled posts are then provided to the data science team for computational system training. The steps of CASM are clearly outlined in the analysis system to help quickly orient each annotator.

Following the methodological and web-based system tutorial, student annotators undergo a week-long immersion on Twitter. This immersion includes passive observation of twenty Twitter users from our corpus to familiarize themselves with the dataset by going through each user's profile, posts, photos, and videos. The annotators are instructed to ethnographically observe the ways users portray themselves online through what they share, who they engage with, and how frequently they post. The Twitter immersion also involves a critical ethical discussion regarding their observation. As a group, student annotators agree to guidelines for protecting the anonymity of users, including: completion of annotations in a private setting, exclusion of users with private accounts, and separation of field notes and identifying information.

After the Twitter immersion, students attend a process meeting to share their observations with other annotators and the expert annotator (the trainer). The meeting is spent training the new annotators to consider contextual features they may be missing from their initial observations. In the second week of training, student annotators annotate 100 Twitter posts. These annotations are thoroughly reviewed by the expert

annotator for any egregious mistakes and patterns of misinterpretation. Some examples of this include misunderstanding various steps of CASM, missing contextual features, and not utilizing web-based resources in the annotation process (e.g., Hipwiki). The expert annotator provides feedback and then the annotators are ready to begin the full annotation process on the official Twitter dataset.

3. Methods for Analysis of Annotator Perspectives

3.1. Qualitative

The postdoctoral research scientist conducted one interview with each domain experts that were employed by the lead author to conduct annotations. The purpose of the interview was to discuss the coding process in general and to review a subset of the annotations in detail to better understand the aspects of images that led to a specific classification. Interviews were conducted at a Chicago-based violence prevention organization in which the domain experts were either employed or affiliated. The mission of the organization is to reduce violence in Chicago by "replacing the cycle of violence using the principles, practices and teachings of nonviolence."

The social science team reviewed two main types of annotation examples. First, we selected examples where a domain expert provided a label that was unique (different from the other domain expert and from the student annotators) across four different classifications: *aggression*, *loss*, *substance use*, or no label. For both of the domain experts we selected 20 unique examples. Second, we selected an additional five examples in each of the four classifications (20 additional examples) where the domain experts agreed with each other, but the social work annotators provided a different label.

The postdoctoral research scientist then conducted separate structured interviews with each domain expert annotator for 30 to 45 minutes. The domain experts described how they interpreted and labeled the tweets. Oral consent was obtained, and both participants were paid an hourly rate for the time it took to conduct the interviews. During the interview, the annotators were asked to describe and explain their responses to 40 tweets with 20 of them overlapping between them. The postdoctoral research scientist reviewed 60 unique tweets in total, which accounts for approximately 10% of the total number of disagreements.

We analyzed the interview data using an inductive qualitative approach. The interviews were transcribed and read on once initially to create a list of preliminary codes. We then applied a codebook to the transcripts and

revised them based on a thorough read. Both transcripts were then coded by two additional authors. We resolved discrepancies through discussion until consensus was achieved. All data was analyzed by hand given the small amount of data. Emerging themes were established by reviewing the transcripts repeatedly and considering the similarities and differences of meaning across the annotators. We will discuss the findings from the interviews with domain experts in Section 4.

3.2. Statistical and computational methods

We compute several statistics for evaluating disagreements between annotators.

Code baselines. First of all, for each annotator (or group of annotators for which we merge their annotations) and code, we compute the overall proportion of positive votes. These proportions will be referred to as code baselines and can be seen as a measure for the annotator’s overall tendency to label a tweet as the respective code. We compute a confidence interval for these numbers (interpreting the decisions as coming from a binomial distribution).

Annotator correlation coefficients. To obtain a general measure of how much disagreement there is in the data, for each class we compute Spearman correlation coefficients for the labels given by two annotators (or group of annotators for which annotations are merged).

Disagreement statistics. For two given annotators (or two groups of annotators for which annotations are merged) and each code, we first calculate the baseline proportion of the number of tweets with conflicting annotations to the overall number of tweets. In addition, for each (textual or visual) concept c we compute the same ratio but only consider tweets that contain the concept c . We compute confidence intervals for the baselines as well as the concept-based ratios as for the code baselines. We use statistical testing to check which concepts significantly affecting disagreement: if the confidence interval for concept c does not overlap with the confidence interval of the respective baseline, this means that for the chosen annotators and code, the concept c has a significant impact on the amount of disagreement between these annotators for this code. Such a difference indicates that the annotators might implicitly assign different code relevance to the respective concept, or, in other words, interpret the concept differently for the task at hand.

Annotator bias. To better understand the reasons for disagreement, for all concepts and codes, we compute the average direction of disagreement. To this end, we first compute differences in code labels for an individual tweet as values -1, 0 and 1 by subtracting the (binary) label of the first annotator from the label given by the

second annotator. We then compute the average and confidence interval for the resulting list of non-zero values over all relevant tweets (i.e. that include the concept of interest). A baseline bias is computed over all tweets and significance is checked similar to the calculation of concept-based disagreement ratios.

Concept disagreement correlations. For each concept and code, we calculate Spearman correlation coefficients between concept presence in the tweets and disagreement in the associated annotations. This provides an additional measure for the importance of individual concepts for disagreement.

Disagreement prediction. We order tweets by annotation times and for different positions x , use the first x tweets for training logistic regression models to predict disagreement with respect to any of the codes, using textual, visual or both kinds of features as model input. All models are then evaluated on the test data which at any time consists of all tweets that have not been used for training. This method has some resemblance to the one proposed in [25] but aims at predicting disagreement instead of the label given by an individual annotator and does not assume the existence of any “true” gold label.

4. Case Study

The corpus for this study comes from the Gang Intervention and Computer Science study, an interdisciplinary project between the SAFElab at the School of Social Work and several members of the Data Science Institute at Columbia University. This project leverages publicly available data from youth and young adults who claim gang association and ties on Twitter and aims to better understand the individual, community, and societal-level factors and conditions that shape aggressive communication online and to determine potential pathways to violence using machine learning.

In order to create our Twitter corpus, we first scraped data from Gakirah Barnes. The first author has studied the Twitter communication of Gakirah Barnes since 2014. Motivations for this study included her age, race, and location, all of which the literature points to as potential risk factors for violence, victimization, and perpetration [23]. Moreover, her assumed gender, physical presentation on Twitter, status within a local Chicago gang, and mentions and subsequent conversations conducted on Twitter regarding two homicides, all made her a unique case study. Gakirah was a 17-year-old female who self-identified as a gang member and “shooter.” After the murder of her close friend Tyquan Tyler, Gakirah changed her Twitter account handle to @TyquanAssassin. Gakirah was

active on Twitter, amassing over 27,000 tweets from December 2011 until her untimely death on April 11, 2014. She used the account to express a range of emotions to include her experiences with love, happiness, trauma, gang violence, and grief.

Our corpus contains 1,851 tweets from 173 unique users scraped in February 2017. Users for this corpus were selected based on their connections with Gakirah Barnes and her top 14 communicators in her Twitter network. Additional users were collected using a snowball sampling technique [2]. For each user we removed all retweets, quote tweets, and tweets without any image, and limited to 20 tweets per user as a strategy to avoid the most active users being overrepresented.

4.1. Qualitative findings

Three themes emerged from the interviews with domain experts, which accounted for the majority of differences between the domain experts and student annotators: recognizing community-level factors, people, and hand gestures.

First, domain experts were able to better recognize community-level factors like places or context. For example, a domain expert identified a handmade card in one of the images. She explained that this type of card was made in and sent from prison. This contextual clue influenced a decision to categorize the photo as *loss*. In another example, a home was featured prominently in a Twitter photo, which had a line of people waiting in front of the house. Both domain experts suggest that this photo presented a house used to distribute illicit drugs. Second, domain experts recognized certain individuals in the Twitter photos. For example, the domain experts reviewed an image with artwork conveying a collection of hand drawn faces. They immediately recognized that each person drawn represented a well-known local rap artist who had been killed. Third, hand gestures in pictures were identified by domain experts as associated with specific gangs and were understood according to the message conveyed. For example, domain experts understood nuanced differences in hand gestures, including the difference between “throwing up” a gang sign (signifying affiliation or association with that gang) versus “throwing down” a gang sign (signifying disrespect towards that gang). In addition to the emergent themes, we also identified challenges with the annotation process. In some instances, domain experts admitted to unintentionally selecting the wrong code, which may reflect the time spent labeling the posts.

¹ For the analysis we exclude two tweets for which we do not have annotations from all annotators.

4.2. Findings from statistical and computational methods

As textual concepts we use the 500 most common words and emojis (computed over all 1,851 tweets), on the visual part we use a list of nine concepts (*handgun, long gun, hand gesture, joint, lean, person, tattoo, marijuana, and money*) which were originally defined for the purpose of training detectors for gang violence prevention and were manually annotated in all images. We run all statistical methods described in Section 3.2, using a confidence value of 0.99 for computing confidence intervals and testing significance.¹

Table 1: Spearman correlation coefficients for psychosocial code annotations from different annotators.

annotators	aggression	loss	substance use
S 1 vs S 2	0.23	0.82	0.75
DE 1 vs DE 2	0.54	0.66	0.73
S vs DE	0.38	0.84	0.78

Annotator correlation coefficients are shown in Table 1. For *loss* and *substance use*, correlations within and between groups are all rather high (0.66 or more), indicating that for these codes, annotators label tweets in a very similar way. However, in case of *aggression* correlation coefficients are much lower. Interestingly, the lowest value of 0.23 was attained for correlation between annotations of the students.

Looking at *annotator baselines* for the different codes (Table 2) reveals that student annotators are in general far less likely to label a tweet as *aggression* as compared to domain experts (2.9% and 4.8% vs 13.4% and 20.3%). This explains how the corresponding correlation coefficient can be much lower for student annotators than for domain experts (0.23 vs 0.54), even though the disagreement baseline is lower for student annotators (5.7% vs 13.4%; see Table 3). For both other codes, baselines for all annotators are much more comparable (see last two columns of Table 2).

These findings point towards general annotator tendencies that provide important insights into the motivations for how Twitter content is labeled. For example, our domain experts may label more content as aggressive as a way to maintain safety in their community. As such, a false negative for aggression is only a minor inconvenience for the student annotators

while a false negative for the domain experts could have lethal consequences for individuals they may know. On the other hand, our student annotators may be biased towards minimizing aggression or other themes that are stereotypical or further marginalized communities of color.

Table 2: Code baselines (including confidence intervals) in percent, of student (S) and domain expert (DE) annotators for labeling tweets as the three psychosocial codes.

annotator/s	aggression	loss	substance use
S 1	2.9 (1.9-3.9)	15.9 (13.7-18.1)	11.7 (9.8-13.6)
S 2	4.8 (3.5-6.1)	15.3 (13.1-17.4)	17.2 (14.9-19.4)
S merged	6.7 (5.2-8.2)	18.0 (15.7-20.3)	12.6 (10.7-14.6)
DE 1	13.4 (11.4-15.4)	18.6 (16.3-20.9)	12.6 (10.7-14.6)
DE 2	20.3 (17.9-22.7)	11.8 (9.9-13.8)	12.3 (10.3-14.2)
DE merged	23.6 (21.0-26.1)	19.9 (17.5-22.3)	15.5 (13.3-17.6)

Table 3 and Table 4 contain *disagreement statistics* for the codes *aggression* and *substance use*. For each feature we state the total number of relevant tweets, the fraction of tweets with conflicting annotations (as difference to the respective baseline), the annotator bias and the Spearman correlation coefficient between concept presence and binary disagreement indicator. The tables only include concepts where the fraction of conflicting annotations was found to be significantly different from the respective baseline.

In the *disagreement statistics* for the code *aggression* (Table 3), for student annotators we can see that *handgun* is the most relevant concept for disagreement (with a correlation coefficient of 0.41), which intuitively makes sense. For disagreements between student annotators and domain experts, the annotator bias of 0.9 shows that irrespective of any concept presence, in 95% of disagreement cases, domain experts voted for *aggression* while student annotators did not. The corresponding correlation coefficient of 0.40 suggests that such disagreements are often related to the presence of *hand gesture* in the image, which is in line with our findings from interviews with domain experts. Additionally, we want to point out that *hand gesture* indicates disagreement between domain experts as well, but this concept was

not found to cause any conflicting annotations between student annotators. In a separate test, it did not significantly increase the likelihood of any student annotator to label a corresponding tweet as *aggression*. This means that without domain expert annotations, the relevance of *hand gesture* to *aggression* would not be visible.

Table 3: Disagreement statistics for the label aggression.

	feature	#tweets	disagr. in %	ann. bias	corr. coeff.
S 1 vs S 2	baseline	1849	5.7	+0.3	-
	(txt) 🗨️	69	+16.0	+0.4	0.14
	(txt) 🖐️	13	+40.4	-0.4	0.15
	(img) <i>handgun</i>	164	+30.9	+0.7	0.41
	(img) <i>long gun</i>	15	+34.3	+1.0	0.13
	DE 1 vs DE 2	baseline	1849	13.4	+0.5
(txt) <i>n***az</i>		13	+40.4	+1.0	0.10
(txt) <i>neva</i>		10	+56.6	+0.7	0.12
(img) <i>hand gesture</i>		572	+15.4	+0.6	0.30
(img) <i>handgun</i>		164	+11.6	+0.6	0.11
S vs DE	baseline	1849	19.0	+0.9	-
	(txt) <i>n***az</i>	13	+50.2	-0.8	0.11
	(txt) <i>neva</i>	10	+51.0	+1.0	0.10
	(img) <i>hand gesture</i>	572	+23.3	+0.9	0.40
	(img) <i>handgun</i>	164	+25.5	+0.9	0.20
	(+6 txt)

Table 4 lists *disagreement statistics* for *substance use*. Here, the presence of *joint* in the image of the tweet correlates with disagreement within both groups and between the two groups (coefficients 0.32, 0.27 and 0.26). For student annotators, there seems to be some additional confusion about the words “dm” and “asl” (+~50% disagreement in presence of each concept) as well as the visual presence of *lean* (+21.3% disagreement). Somewhat surprising is the finding that

handgun increases the chance of conflict between student annotators and domain experts for the label *substance use*.

Table 4: Disagreement statistics for *substance use*. All concepts with statistically significant differences to the respective baseline are included.

	feature	#tweets	disagr. in %	ann. bias	corr. coeff.
S 1 vs S 2	baseline	1849	6.6	0.8	-
	(txt) <i>dm</i>	9	+49.0	+1.0	0.14
	(txt) <i>asl</i>	7	+50.5	+1.0	0.13
	(img) <i>lean</i>	43	+21.3	+0.8	0.13
	(img) <i>joint</i>	185	+23.7	+0.9	0.32
DE 1 vs DE 2	baseline	1849	6.0	-0.1	-
	(img) <i>joint</i>	185	+19.4	+0.2	0.27
S vs DE	baseline	1849	6.2	-0.4	-
	(img) <i>joint</i>	185	+18.7	-0.9	0.26
	(img) <i>handgun</i>	164	+10.3	-0.9	0.13

Check-in’s with student annotators revealed a disparate meaning-making process. For example, “dm” or direct messaging may trigger for a student annotator questions about the types of conversations that happen during a private exchange. At times the annotators misunderstood the phonetic interpretation of “asl” which in the context of our study would be used to phonetically spell a word like “as hell”. The presence of a Styrofoam cup would trigger a label of an entire tweet as “lean” whereas another student annotator would not identify the entire tweet as *substance use*. Lastly, the socio-political interpretation of what a *handgun* means in an image with young African American youth informed how the student annotators labeled *substance use*.

Annotator bias. The only case where the presence of a concept significantly alters the bias for disagreement is in case of code *substance use* and visual concept *joint* for student vs domain expert disagreement. Apparently, in almost all cases (-0.9 annotator bias, i.e. around 95%) of *substance use* disagreement with a joint present in the image, student annotators voted for *substance use* and domain experts did not (as compared to the concept-independent baseline bias of around 70%). This suggests that student annotators saw *joint* as far more indicative for *substance use* than domain experts.

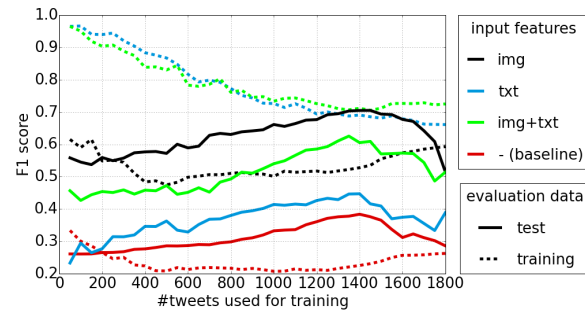


Figure 1: Performances of logistic regression models predicting disagreement between S and DE annotators for any code.

Figure 1 shows F1 scores from our experiments on predicting disagreements between student annotators and domain experts, comparing models that use visual, textual or both types of features. Since tweets are ordered by annotation time for this experiment, the plot visualizes the development of performances over the course of the annotation process, where at any point all current annotations are used for training and all future annotations are used as test set.

As a statistical baseline we also include performances of a system that knows the true ratio $p\%$ of items with disagreement in the evaluation data and (deterministically) classifies $p\%$ of the tweets with disagreement and $p\%$ of the tweets without disagreement as having disagreement. Note that the F1 score of this baseline is given by $p/100$, hence it directly describes the ratio of tweets with disagreement in the data set.

In the plot we see that, using only visual features, already after 50 tweets the prediction model achieves an F1 score of around 0.55, which is far above the respective baseline of around 0.25. For the most part, this difference remains nearly constant. The drop of performance at the end is likely due to the small number of remaining tweets for testing.

We find that for our data, adding textual concepts is detrimental to performance on unseen data, where the visual model consistently outperforms both other models and using text alone gives the worst results. Using only textual features still leads to above-baseline prediction if more than 200 tweets are used for training, but this difference remains comparatively small until the end. Considering performances on the training data clearly shows that whenever textual concepts are used as input features, prediction models apparently learn to use noise in the training set for prediction and thereby fail to generalize to the test data, a typical case of overfitting. However, this effect is getting smaller as more tweets become available for training, especially for the model that uses both visual and textual features.

Also note that the textual features we used for the experiment are more low-level and higher in magnitude as compared to our visual features. Therefore, the text modality should not be deemed generally useless for disagreement prediction based on these results.

5. Discussion

In this paper, we examine disagreements between domain experts and student annotators to consider the promise and challenge of annotating Twitter data from vulnerable and marginalized communities. Leveraging annotations from a Chicago Twitter corpus of African American and Latino youth and young adults who live in communities with high rates of gang violence, we underscore the importance of considering annotator background and involving local domain experts when analyzing Twitter data for machine learning. Specifically, nuances in culture, language, and local concepts should be considered when identifying annotators and the type of training they should receive before reviewing Twitter data. Furthermore, our findings emphasize the importance of identifying interpretation-related problems in annotation and the need for strategies on treating disagreement based on its causes.

5.1. Annotation conflicts

Much of the computer science literature focuses on eliminating disagreements between annotators, but here we argue that in the case of data from marginalized populations, some disagreement may not be negative. As we have seen, even if it is doubtful whether there really is an objective “gold standard” for the final labels, analyzing disagreements can lead to a better understanding of the domain of application. Especially in this context of more complex use-cases, if annotations are done by a few trained annotators, one can monitor their annotations and discuss disagreements as they arise, successively leading to higher quality of the annotations and a more complete overall picture.

By comparing disagreements between and within two groups of annotators, domain experts and student annotators, we uncovered critical differences in interpretation of behaviors in images on Twitter. Symbolic Interactionism theory suggests that individuals use gestures - “any part of an ongoing action that signifies the larger act or meaning” (pp. 9) to understand human behavior [5]. For example, a domain expert who lives in the same or similar community as the Twitter users under study would have a nuanced understanding of the use of the gun emoji or a specific hang gesture. They are able to situate what those

specific gestures meaning within the local context, thus informing if the gesture should be determined threatening.

When gestures are interpreted incorrectly, we risk inflicting a detrimental and compounded impact on the current and future experiences for marginalized users already experiencing the results of systematic oppression. Patton et al. [16] uncovered distinct differences in how police use social media as evidence in criminal proceeding. For example, the misinterpretation of gestures made by young African American men on Facebook led to the arrest of over 100 young Black men in New York City, some of whom were not involved with the crime under question [22]. Conversely, social media threats made by a White male, Dylann Roof, who killed nine African American church-goers in Charleston, South Carolina, went undetected by law enforcement. In addition, Safiya Noble [11] warns us that biases unchecked in the labeling of images on google reinforce racist stereotypes.

Understanding and analyzing disagreements benefitted our annotators. At the micro level, this process pushed our student annotators to redefine labels that could lead toward providing a user with additional supports and resources. At the macro level our processes forced us to consider how applying the wrong label could further criminalize an already stigmatized population. For example, interpreting a *hand gesture* that represents gang association in case of *aggression* only became evident after consulting with experts, so the “true” meaning of *hand gesture* would have been missed by our student annotators. This implies that the common strategy of adding more non-expert annotators would likely not have revealed this aspect either.

Luckily, we found that computational models can learn to predict disagreement between social work annotators and domain experts from rather few samples when using suitable features for the prediction. In practice this can potentially be useful for better leveraging community members’ expertise by automatically selecting critical examples for expert annotation. Essentially, this would mean adopting an active learning paradigm for selectively collecting annotations, similar to [24], but instead of focusing on detectors, expert annotations would be selected in order to train annotators or content moderators.

5.2. Role of domain experts in annotation

Domain expertise is vital to annotating Twitter data from marginalized communities. In the study of gang-related concepts on Twitter, we hired domain experts to perform several functions. First, we leveraged insights from domain experts to train student annotators on

nuances in language, culture and context that are embedded in the text and images in the Twitter posts. Second, domain experts separately annotated Twitter posts from users in their own community, which allowed us to compare their annotations to graduate student annotations. These annotations help us understand how people from the community naturally view Twitter posts using their experience and expertise. Third, we interviewed them to understand how they made decisions and what informed the labels they assigned to images. Interviews with the domain experts revealed critical concepts like handmade cards or recognizing people which were visible in the images, but not captured by our visual concepts. The critical concepts are not frequent and thus challenging to detect using statistical or automatic methods. Even if it were possible to detect these concepts it would be impossible to find out the extent to which a hand gesture is important without interviewing the domain experts.

Domain experts and student annotators engage the annotation process differently. Our domain experts have more intuitive and instinctive interpretations of Twitter posts because those posts reflect their everyday lived experiences and how they interpret their community. Conversely, the student annotators are trained to annotate using a detailed process, specifically considering alternative meanings and interpretations because they do not have the same contextual experiences. Weighing the differences between domain experts and student annotators should be informed by the research question and specific tasks. In this study, domain experts provide a nuanced understanding of language and behavior (e.g. hand gestures) that our student annotators would only understand if they had the same lived experiences. Our student annotators pushed us to consider the broader ethical challenges that come with annotating Twitter data from African American and Latino youth and young adults.

5.3. Ethical considerations

As researchers who study gang-related content on Twitter, we understand our ethical obligations to ensure that our work does not further criminalize or stigmatize African American users in our sample. To protect the users in our corpus, we will only publish specific parts of the statistical features to prevent the ability to trace our users. Given the popular use of social media to monitor communities for potential acts of violence, this study underscores the importance of domain expertise and studying disagreement to highlight challenges in perception and interpretation of Twitter data from marginalized communities.

6. Future Directions

This work has implications for the development of and training for content moderation at social media platforms. Companies like Facebook and Twitter might consider training sessions where disagreements between moderators are identified and reviewed to identify moderator bias and gain additional contextual and cultural insights that may inform how they make decisions about removing content.

As another step, we plan to apply our methods for annotator perspective analysis in several other scenarios. First, we plan to use annotations from different datasets, such as text-only tweets of gang-involved youth [4] or even annotations of image captions on Flickr collected over crowdsourcing platforms [3]. Second, we want to test how generalizable these methods are by using them to evaluate misclassifications of machine learning algorithms, which can be seen as disagreement between a detector and human annotators, or to compare functioning of multiple automatic methods.

References

- [1] Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163-222). Springer, Boston, MA.
- [2] Atkinson, R., & Flint, J. (2001). Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social research update*, 33(1), 1-4.
- [3] Blandfort, P., Karayil, T., Borth, D., & Dengel, A. (2017, October). Image Captioning in the Wild: How People Caption Images on Flickr. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes* (pp. 21-29). ACM.
- [4] Blevins, T., Kwiatkowski, R., Macbeth, J., McKeown, K., Patton, D., & Rambow, O. (2016). Automatically processing tweets from gang-involved youth: towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2196-2206).
- [5] Blummer, H. (1969). *Symbolic Interactionism. Perspective and Method*. University of California Press. Berkeley, CA.
- [6] Brock, A. (2018). Critical technocultural discourse analysis. *New Media & Society*, 20(3), 1012-1030.
- [7] Broussard, M. When Cops Check Facebook. *The Atlantic*, 2018.
<https://www.theatlantic.com/politics/archive/2015/04/when-cops-check-facebook/390882/>.

- [8] Frey, W. R., Patton, D. U., Gaskell, M. B., & McGregor, K. A. (2018). Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data. *Social Science Computer Review*, 0894439318788314.
- [9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
- [10] Miltisakaki, E., Joshi, A., Prasad, R., & Webber, B. (2004). Annotating discourse connectives and their arguments. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*.
- [11] Noble, Safiya. (2018) Algorithms of Oppression. How Search Engines Reinforce Racism. *New York University Press*. New York.
- [12] Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1), 71-106.
- [13] Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. *Language Resources and Evaluation*.
- [14] Passonneau, R. J. (2004, May). Computing Reliability for Coreference Annotation. In *LREC*.
- [15] Patton, D.U., Blandfort, P., Frey, W.R., Schifanella, R., & McGregor, K. (under review). VATAS: An open-source web platform for visual and textual analysis of social media.
- [16] Patton, D. U., Brunton, D. W., Dixon, A., Miller, R. J., Leonard, P., & Hackman, R. (2017). Stop and Frisk Online: Theorizing Everyday Racism in Digital Policing in the Use of Social Media for Identification of Criminal Conduct and Associations. *Social Media + Society*, 3(3), 2056305117733344.
- [17] Patton, D.U., Frey, W.R., McGregor, K.A., Lee, F.T., McKeown, K.R. (under review). Contextual analysis of social media: a qualitative approach to eliciting context in social media posts with natural language processing.
- [18] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11(Apr), 1297-1322.
- [19] Reidsma, D., & Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, 34(3), 319-326.
- [20] Roberts, S.T. (2016). Commercial content moderation: Digital laborers' dirty work. In Noble, S.U. and Tynes, B. (Eds.), *The intersectional internet: Race, sex, class and culture online* (2016), 147-159. Download here: <https://illusionofvolition.com/publications-and-media/>
- [21] Rodrigues, F., Pereira, F., & Ribeiro, B. (2013). Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12), 1428-1436.
- [22] Speri, Alison (2014). The Kids Arrested in the Largest Gang Bust in New York City History Got Caught because of Facebook. *VICE*. Retrieved here: <https://news.vice.com/>
- [23] University of Chicago Crime Lab (2016). Gun Violence in 2016. Chicago, IL. Retrieved from: <http://urbanlabs.uchicago.edu/attachments/store/2435a5d4658e2ca19f4f225b810ce0dbdb9231cbdb8d702e784087469ee3/UChicagoCrimeLab+Gun+Violence+in+Chicago+2016.pdf>
- [24] Yan, Y., Rosales, R., Fung, G., & Dy, J. G. (2011, June). Active learning from crowds. In *ICML* (Vol. 11, pp. 1161-1168).
- [25] Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., ... & Dy, J. (2010, March). Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 932-939).

Chapter 7

Conclusion

We first summarize the outcomes of our studies on the research questions (Q1.1), (Q1.2) and (Q1.3) (see Section 1.3), before we draw conclusions for the main research question (see Section 1.1).

7.1 Modeling Interpretation

7.1.1 Summary

We introduced a unifying theoretical framework for modeling interpretation that is applicable to interpretation of both human and computer systems. This framework revealed connections between various fields (statistics, pattern mining, visualization, deep learning) and possibilities for computational analysis of interpretation. In our survey we mention many well-known approaches for analysis but also explain less common possibilities for comparing several ways of interpretation with computational methods. We explored one such possibility with movie ratings data (pages 95ff.). There, our theoretic framework enabled us to propose an evaluation of learned vector representations for the users' ways of interpretation. Interestingly, we found that the default way of combining user and movie information (concatenation) gave the best user representations but was inferior to other fusion strategies in terms of prediction ability.

7.1.2 Future Work

Our paper on user embeddings (pages 95ff.) can easily be extended by considering further fusion strategies and varying hyper-parameters of the evaluation metric to examine in more detail how the learned embedding spaces are structured. Doing so would constitute one further step toward a better understanding of deep learning for interpretation analysis, which would be beneficial to address open issues such as the reliability issue we found during our survey (pages 39ff.): How reliable are analysis results coming from complex model-based approaches such as neural networks? Another useful study in this direction could be to run simulations with perspectives that are described by known functions (and have known properties) and then check which analysis models are able to recover which of the properties after fitting the data.

In any case, our efforts on investigating the basics of computational approaches to interpretation not only revealed problems to be fixed but showed new opportunities as well. Importantly, introducing the bearer (i.e., entity that is interpreting) into our modeling lead to a mathematical formulation which can be used for comparing perspectives as well as analyzing individual ones.

In particular, we pointed out that our problem formulation goes analogously to the one in the mathematical branch of Functional Data Analysis (pages 95ff.). This relation explains how distances between ways of interpretation can mathematically be understood as distances between functions and calculated as such. We have seen in the same paper how this insight can be used to cluster users based on their ways of interpretation. This approach can easily be extended to cluster other contextual influences (e.g., geolocation of author, time of the day, weather) in terms of their effects on interpretation. More generally, it would be very interesting to analyze various contextual effects (such as the ones mentioned in Section 2.3) using fusion models for analysis. This would not only allow to quantify the overall effect (e.g., in terms of performance boost you get by adding the factor as context), but one might find more complex interactions between input features and contextual influences by using analysis techniques like heatmapping.

7.2 Subjective Image Interpretation

7.2.1 Summary

In our crowdsourcing study we found that subjectivity is important to consider if the goal is to caption images in a natural way. Within the same study, we structured the image captioning space in terms of subjectivity, visibility and purpose of captions by means of manual annotation. This led to an annotated dataset which we released and believe to be useful for guiding further progress on captioning images in a more natural way.

As a first step into this direction, we proposed an interpretable and intuitive method for generating affective image captions. We also proposed the FAV model for subjective image interpretation in terms of concepts, which allows for a more structured prediction of subjective image interpretation and enabled us to train and evaluate models for predicting unseen adjective-noun combinations. Our corresponding dataset *aspects-DB* was made publicly available as well.

7.2.2 Future Work

As next steps, two extensions of this work could be considered: First, aspect-based subjective interpretation (instead of VSO) can be used as concept extraction for subjective image captioning. Second, additional context-signals such as user ID could be incorporated into the prediction models for predicting context-dependent subjective interpretation. Given additional data of the form (image,user,noun,aspect,value), this could be achieved by using user information as additional context signal (together with noun information).

7.3 Gang Violence Prevention

7.3.1 Summary

We introduced a multimodal model for detecting the psychosocial codes *aggression*, *loss* and *substance use* from tweets of presumably gang-involved youth. By adding visual information to the detection, we were able to achieve a large performance gain as compared to text-only detection models. Interestingly, we found that for different codes different modalities worked best, suggesting that visual and textual information is indeed complementary in some cases. On the visual side, we also introduced a list of visual concepts of interest, for which we built an individual detection model. These visual concepts proved useful for analyzing annotator disagreements and increase the

explainability of our multimodal detection model. During the process we built a new dataset which we share with other researchers (under strict conditions due to the delicate nature of the data).

Our dataset and the detection methods were built in a collaborative way together with social work researchers. This collaboration was highly beneficial for getting better knowledge and avoiding mistakes. For example, the drug lean (a drug made of soda and cough syrup) would have been missed by me since I was not aware of it before discussing with our social work team. We provide a blueprint for such an interdisciplinary collaboration for high-quality annotation and make our custom annotation platform VATAS available as open-source. Furthermore, we developed methods for analyzing annotator disagreement and saw how analyzing annotation conflicts can indeed help to improve understanding of a specific domain. In our specific case, disagreement analysis revealed some nuances in meanings we were not aware of before, such as the different meaning of hand gesture depending on where it is pointed, which would have been missed in a purely data-driven computer science approach.

7.3.2 Future Work

The work on gang violence prevention we described in this dissertation is part of a long-term project which is still at an early stage. In particular, our detection system is not yet deployed in the wild, and not all of our annotations are used yet. However, the results already look promising and the plan is to check how useful our methods are in the field. Currently, our dataset is being analyzed for understanding dynamics of codes.

7.4 Main Research Question

Zooming out to the main research question (Q1) of how we can enable computers to interpret multimedia messages in a subjective way: In two application studies we have seen that this can generally be achieved by obtaining labeled samples of multimedia messages and then training a machine learning model to fit this data. So in short, a standard supervised learning approach can be used, as long as suitable data is available and the task is modeled appropriately. How well results will be depends on several factors such as the quality and quantity of the training data, the chosen model architecture, and how exactly the task is modeled.

In both application studies, we built our own datasets (aspects-DB for image interpretation, annotated tweets of gang-involved youth) and informed the modeling process by manual analysis (crowdsourcing for images, collaboration with social work researchers for gang domain). The amount of work we had to put into these points suggests that as powerful as current machine learning models might be, human intelligence and intuition are still necessary for modeling and collecting datasets, and therefore remain crucial in the process of developing intelligent computer systems.

7.4.1 Domain-specific Findings

Furthermore, we found domain-specific details that can become relevant for predicting subjective interpretation.

Our work on subjective image interpretation has shown several points that are relevant for other applications as well: User-data such as user-generated image titles or tags can in principle be used to train automatic detection methods. Still, such data generally has to be cleaned up as users often post irrelevant or otherwise misleading information (“noise” or “spam”). Another part that was particularly important for

this application study was the modeling part, where we found that working with a very general purpose of detection poses an extra challenge on deciding which concepts are suitable to detect.

Regarding the gang violence study, we described a general approach for building multimodal analysis systems (outlined in Section 5.3) which is directly transferrable to other application domains. Moreover, this application study showed how complex and important annotation can be. Due to the difficulty of obtaining ground truth labels, but also due to strong ethical implications of the study, the focus was naturally moved toward annotation and domain understanding. Here, our approach of interdisciplinary collaboration between computer science and social work for annotation and analyzing disagreements was clearly beneficial for avoiding misinterpretations, collecting high quality annotations and developing a deeper understanding of the domain at hand.

7.4.2 Ethics

During this dissertation, some ethical points became relevant while working on gang violence prevention, which have implications beyond applications in this particular area.

In a domain such as gang-associated youth, misinterpreting social media posts can have detrimental consequences. For example, resources for intervention might be assigned inefficiently, innocent people might be arrested, or the trust in law enforcement within the community might further decrease as a result. Such adverse effects can come from building tools with poor quality (which in turn can come from bad datasets), but also from building certain kinds of tools in the first place.

Moreover, in computer science research it is generally expected that datasets are shared so that results can be reproduced and methods further improved, but publishing could cause harm to involved individuals, especially if annotations are with respect to “negative” concepts related to drugs and violence.

We took several steps to address these points, such as not including any actual tweets or images of gang-associated youth into presentations or publications (including this dissertation)¹ and working together closely with social work researchers (who also communicated with domain experts from Chicago to prevent problematic misinterpretations). More generally, we realized that doing research on people’s data should be expected to affect their private lives and believe that similar interdisciplinary collaborations will become more and more necessary to advance the field in an ethical way.

7.4.3 Overcoming Challenges

Despite all efforts, predicting subjective interpretation remains a challenging task and performance cannot be expected to be perfect. This was one of the reasons why we incorporated task-specific mid-level concepts into several of our models (adjective-noun pairs for subjective image captioning, visual concepts for psychosocial code detection). Such mid-level concepts can be used to analyze disagreement and therefore help to understand subjective interpretation, as we have seen in case of gang violence prevention, where we used visual concepts for analyzing annotation conflicts. Furthermore, performance on mid-level concepts can be tested separately during development, and during deployment of the final model the detected mid-level concepts serve as explanation for the models decision. For these reasons, we can also expect mid-level concepts to be helpful when dealing with subjective interpretation for application domains different from the ones discussed in this dissertation.

¹even though the tweets we used were all publicly available

Related, we reviewed possibilities for analyzing how trained models make their decisions, which will surely be beneficial to guide further progress. In particular, we still lack a good understanding of what makes one neural network architecture perform much better on a task than another. Comparing neural network models using the reviewed analysis techniques might contribute to such an understanding which would allow for modifying architectures in a more directed way.

7.5 Overall Lessons

Working on this dissertation topic taught me a number of valuable lessons:

- Having suitable datasets is very important for training prediction models, but at the same time, the process of building datasets often is a tedious one if high quality is a criterion.
- Sometimes simple models can be very effective. For example in our work on aspect detection (pages 103ff.), the logistic regression baselines (applied on neural network features) turned out to be surprisingly competitive. In fact, in a previous version of our FAV paper, logistic regression was among the best performing models. So it is important to include meaningful baselines in experiments on prediction, and check whether it is really necessary to use any highly complex prediction model.
- Neural networks are powerful but working with neural networks takes time if one deviates from the highly-optimized standard architectures. In some of my own experiments with custom neural network architectures I had to spend a long time on optimizing hyper-parameters to even outperform baselines such as logistic regression. On the other hand, if the network is optimized properly it often gives very good results.

One last point I want to stress is that, to me, research largely means team work. Most parts of the work presented in this dissertation were done with colleagues from various institutions and of various academic backgrounds. These collaborations helped a lot to improve the quality of results, to acquire new skills and knowledge, and also to become more aware of my own biases in thinking. In particular I would like to point out that my collaborations with researchers from other fields were very insightful. Such collaborations might take some extra effort in the beginning for getting used to work with researchers from other fields, but in my opinion the added value is well worth the extra step.

Acknowledgments

For most of my PhD studies, I received a stipend from the Graduate School of the Center for Cognitive Science at TU Kaiserslautern, where the Psycholinguistics Group at TU Kaiserslautern and the German Research Center for Artificial Intelligence (DFKI) both kindly contributed. After three years, this stipend was extended by the computer science faculty at TU Kaiserslautern. DFKI was also covering most of my traveling expenses, which was very helpful for distributing my research ideas and establishing new connections without becoming overly concerned about being able to still buy food or pay the rent. In addition, I was lucky enough to receive another stipend from DAAD for financing some of my research stay at Columbia University.

Furthermore, a big thank you goes to all my collaborators, not only for doing a significant part of the work presented above, but also for being a steady source of motivation, providing feedback and broadening my own perspective. Tushar Karayil shall be mentioned specifically at this point, as we worked together very closely throughout my dissertation, shared plenty of great conversations, and he has always been supportive. I thank all other members of the Smart Data and Knowledge Services group (especially the individuals who make the MADM group) at DFKI, the Psycholinguistics Group at TU Kaiserslautern and the DVMM Lab at Columbia University for the many interesting discussions, which enriched my own thinking and certainly helped my research. For walking the final few steps, my colleagues from the MADM group further assisted me by sharing extremely helpful comments about the test talk for my disputation and taking on some of the organizational tasks, which I both highly appreciate. Something that is easily taken for granted but is absolutely necessary for most research in machine learning is the technical infrastructure. Luckily, the ISG team at DFKI took great care of this part. In particular, I would like to thank Dr. Christian Schulze for all the technical support and managing the machines used in my experiments.

Even a dissertation requires a certain amount of guidance. In this regard, I would like to thank several more experienced researchers who gave me direct feedback on my work and progress: At the early stages, Prof. Dr. Damian Borth provided much assistance, shared many valuable insights and initiated my research stay at Columbia University with Prof. Shih-Fu Chang. Prof. Chang was an excellent host and greatly helped to improve the quality of my research by demonstrating great rigorousness and discussing with me about many details of my work. He also got me in touch with many other researchers, including Prof. Desmond U. Patton, who became a valued collaborator and inspired me to look at the bigger scope and complexity of online communication. At the later stages of my dissertation, Dr. Jörn Hees assisted me by giving plenty of tips and feedback, in particular about organizing and communicating my research. It is safe to say that without him, the writing and presentation of my thesis would have remained less clear, and the process of finishing it would have been more stressful than it already was. Finally, I thank my advisors Prof. Dr. Shanley E. M. Allen and Prof. Dr. Prof. h.c. Andreas Dengel for enabling me to work on this thesis topic in the first place and accompanying me the whole way. Andreas played a big role in making the dissertation possible, by providing me with a working place with connection to the DFKI infrastructure (and close proximity to a coffee machine). Moreover, I appreciate the useful pointers he gave me in our discussions and the unique perspective he had to offer. Shanley showed a lot of patience and empathy in the many discussions we had, and always took enough time to keep me on track by helping me to think things through and make appropriate decisions. The fact that my work turned out to be only marginally related to linguistics makes me appreciate this even more.

Last but not least, I would like to express my gratitude to all people who are personally close to me. This special group of people moved me when I was stuck in a

place, helped me to experience joy and meaning, and showed me how important and beautiful subjectivity can be. After all, without subjectivity, how would one even define beauty or importance?

About the Author

Philipp Blandfort started studying mathematics at TU Kaiserslautern in October 2009. He received his Bachelor of Science in March 2013 and his Master of Science in April 2015, both from the department of mathematics. In January 2016, Blandfort began his dissertation studies as a PhD candidate in Cognitive Science at TU Kaiserslautern in collaboration with the German Research Center for Artificial Intelligence (DFKI). While working on his dissertation, he visited the Digital Video and Multimedia Lab at Columbia University from November 2016 to March 2017. Since October 2019, Blandfort makes his living by freelancing as IT consultant.

Bibliography

- [1] Margarita Vázquez Campos and Antonio Manuel Liz Gutiérrez. The notion of point of view. In *Temporal Points of View*, pages 1–57. Springer, 2015.
- [2] Jens Zimmermann. *Hermeneutics: A very short introduction*. OUP Oxford, 2015.
- [3] Simon Watts and Paul Stenner. Doing q methodology: theory, method and interpretation. *Qualitative research in psychology*, 2(1):67–91, 2005.
- [4] Heather Tan, Anne Wilson, and Ian Olver. Ricoeur’s theory of interpretation: An instrument for data interpretation in hermeneutic phenomenology. *International Journal of Qualitative Methods*, 8(4):1–15, 2009.
- [5] Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):30:31–30:57, June 2018.
- [6] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1 – 15, 2018.
- [7] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM ’13, pages 223–232, New York, NY, USA, 2013. ACM.
- [8] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.
- [9] Brendan Jou and Shih-Fu Chang. Deep cross residual learning for multitask visual recognition. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 998–1007. ACM, 2016.
- [10] Terra Blevins, Robert Kwiatkowski, Jamie Macbeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. Automatically processing tweets from gang-involved youth: towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206, 2016.
- [11] GlobalWebIndex. *Flagship Report 2018*, 2018. <https://www.globalwebindex.com/hubfs/Downloads/Social-H2-2018-report.pdf>. Accessed May 2019.
- [12] Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, Mei-Ling Shyu, and SS Iyengar. Multimedia big data analytics: A survey. *ACM Computing Surveys (CSUR)*, 51(1):10, 2018.

- [13] Belle Beth Cooper. *How Twitter's Expanded Images Increase Clicks, Retweets and Favorites*. buffer, 2013. <http://disq.us/t/wgoiyz>. Accessed May 2019.
- [14] Sameer Hinduja and Justin W Patchin. Offline consequences of online victimization: School violence and delinquency. *Journal of school violence*, 6(3):89–112, 2007.
- [15] Jonathan E King, Carolyn E Walpole, and Kristi Lamon. Surf and turf wars online—growing implications of internet gang violence. *Journal of Adolescent Health*, 41(6):S66–S68, 2007.
- [16] Junghyun Kim, Robert LaRose, and Wei Peng. Loneliness as the cause and the effect of problematic internet use: The relationship between internet use and psychological well-being. *CyberPsychology & Behavior*, 12(4):451–455, 2009.
- [17] VA Shiva Ayyadurai. System and method for content-sensitive automatic reply message generation for text-based asynchronous communications, April 6 2004. US Patent 6,718,368.
- [18] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201. ACM, 2010.
- [19] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1185–1194. ACM, 2010.
- [20] Kuo-Yen Lo, Keng-Hao Liu, and Chu-Song Chen. Assessment of photo aesthetics with efficiency. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2186–2189. IEEE, 2012.
- [21] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.
- [22] Max Kapustin, Jens Ludwig, Marc Punkay, Kimberley Smith, Lauren Spiegel, and David Welgus. Gun violence in chicago, 2016. *Univ of Chicago Crime Lab*, 2017.
- [23] Desmond Upton Patton, Robert D Eschmann, and Dirk A Butler. Internet banging: New trends in social media, gang violence, masculinity and hip hop. *Computers in Human Behavior*, 29(5):A54–A59, 2013.
- [24] Jeffrey Lane. The digital street: An ethnographic study of networked street life in harlem. *American Behavioral Scientist*, 60(1):43–58, 2016.
- [25] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, 1(1):39–48, 2018.
- [26] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

- [27] Merriam-Webster. *Definition of "Interpretation"*. <https://www.merriam-webster.com/dictionary/interpretation>. Accessed April 2019.
- [28] Merriam-Webster. *Definition of "to interpret"*. <https://www.merriam-webster.com/dictionary/interpret>. Accessed April 2019.
- [29] Oxford Dictionaries. *Definition of "to interpret"*. <https://en.oxforddictionaries.com/definition/interpret>. Accessed April 2019.
- [30] Cambridge English Dictionary. *Definition of "to interpret"*. <https://dictionary.cambridge.org/dictionary/english/interpret>. Accessed April 2019.
- [31] Martin Kusch. Epistemischer relativismus. In *Handbuch Erkenntnistheorie*, pages 338–344. Springer, 2019.
- [32] Alan Cruse. *Meaning in language: An introduction to semantics and pragmatics*. Oxford University Press, 2011.
- [33] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [34] Geoffrey N Leech. *Principles of pragmatics*. Routledge, 1983.
- [35] H Paul Grice, Peter Cole, Jerry L Morgan, et al. Logic and conversation. 1975, pages 41–58, 1975.
- [36] Victor Lavrenko, Raghavan Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *Advances in neural information processing systems*, pages 553–560, 2004.
- [37] Frank Keller. Jointly representing images and text: Dependency graphs, word senses, and multimodal embeddings. In *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion*, pages 35–36. ACM, 2016.
- [38] Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.
- [39] Marco Baroni and Gemma Boleda. *Slides for the course "Distributional Semantics"*. <https://www.cs.utexas.edu/~mooney/cs388/slides/dist-sem-intro-NLP-class-UT.pdf>. Accessed May 2019.
- [40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [41] Lesley Jeffries and Daniel McIntyre. *Stylistics*. Cambridge University Press, 2010.
- [42] P Ivan Pavlov. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, 17(3):136, 2010.
- [43] Martin J Pickering and Holly P Branigan. Syntactic priming in language production. *Trends in cognitive sciences*, 3(4):136–141, 1999.
- [44] Roberto Dell’Acqua and Jonathan Grainger. Unconscious semantic priming from pictures. *Cognition*, 73(1):B1–B15, 1999.

- [45] John A Bargh, Mark Chen, and Lara Burrows. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of personality and social psychology*, 71(2):230, 1996.
- [46] Buster Benson. *Cognitive bias cheat sheet*. Better Humans, 2016. <https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18>. Accessed April 2019.
- [47] Wikipedia. *List of cognitive biases*. https://en.wikipedia.org/wiki/List_of_cognitive_biases. Accessed April 2019.
- [48] Daniel Kahneman and Patrick Egan. *Thinking, fast and slow*, volume 1. Farrar, Straus and Giroux New York, 2011.
- [49] Yair Bar-Haim, Dominique Lamy, Lee Pergamin, Marian J Bakermans-Kranenburg, and Marinus H Van Ijzendoorn. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. *Psychological bulletin*, 133(1):1, 2007.
- [50] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
- [51] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.
- [52] Lee Ross. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology*, volume 10, pages 173–220. Elsevier, 1977.
- [53] Scott T Allison and David M Messick. The group attribution error. *Journal of Experimental Social Psychology*, 21(6):563–579, 1985.
- [54] Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 162–170. ACM, 2018.
- [55] Stanley Milgram. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371, 1963.
- [56] Allen L Edwards. The social desirability variable in personality assessment and research. *Journal of the American Statistical Association*, 1957.
- [57] Anton J Nederhof. Methods of coping with social desirability bias: A review. *European journal of social psychology*, 15(3):263–280, 1985.
- [58] Cindy Gallois and Howard Giles. Communication accommodation theory. *The international encyclopedia of language and social interaction*, pages 1–18, 2015.
- [59] Owen Hargie. *Skilled interpersonal communication: Research, theory and practice*. Routledge, 2016.
- [60] Lewis Donohew, Howard E Sypher, and E Tory Higgins. *Communication, Social Cognition, and Affect (PLE: Emotion)*. Psychology Press, 2015.
- [61] Jack K Chambers, Peter Trudgill, and Natalie Schilling-Estes. *The handbook of language variation and change*. Wiley Online Library, 2002.
- [62] Merriam-Webster. *Definition of “Preference”*. <https://www.merriam-webster.com/dictionary/preference>. Accessed April 2019.

- [63] Peter Muris and Andy P Field. Distorted cognition and pathological anxiety in children and adolescents. *Cognition and emotion*, 22(3):395–421, 2008.
- [64] Petra Helmond, Geertjan Overbeek, Daniel Brugman, and John C Gibbs. A meta-analysis on cognitive distortions and externalizing problem behavior: Associations, moderators, and treatment effectiveness. *Criminal justice and behavior*, 42(3):245–262, 2015.
- [65] Antonio R Damasio. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 351(1346):1413–1420, 1996.
- [66] Antoine Bechara, Hanna Damasio, and Antonio R Damasio. Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295–307, 2000.
- [67] Antonio Manuel Liz Gutiérrez and Margarita Vázquez Campos. Subjective and objective aspects of points of view. In *Temporal Points of View*, pages 59–104. Springer, 2015.
- [68] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press, 2011.
- [69] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.
- [70] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.
- [71] Federico Raue, Sebastian Palacio, Andreas Dengel, and Marcus Liwicki. Class-less association using neural networks. In *International Conference on Artificial Neural Networks*, pages 165–173. Springer, 2017.
- [72] Stephan Trenn. Multilayer perceptrons: Approximation order and necessary number of hidden units. *IEEE transactions on neural networks*, 19(5):836–844, 2008.
- [73] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [74] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [75] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
- [76] Fang Liu, Clement Yu, and Weiyi Meng. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on knowledge and data engineering*, 16(1):28–40, 2004.
- [77] Alan Hanjalic, Christoph Kofler, and Martha Larson. Intent and its discontents: the user at the wheel of the online video search engine. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1239–1248. ACM, 2012.
- [78] Peter W Foltz and Susan T Dumais. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, 35(12):51–60, 1992.

- [79] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin. *COMMUNICATIONS OF THE ACM*, 35(12):29–38, 1992.
- [80] Ritendra Datta, Jia Li, and James Z Wang. Content-based image retrieval: approaches and trends of the new age. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 253–262. ACM, 2005.
- [81] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.
- [82] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015.
- [83] Tun Thura Thet, Jin-Cheon Na, and Christopher SG Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848, 2010.
- [84] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, 2015.
- [85] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30, 2016.
- [86] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Federica Bisio. Sentic lda: Improving on lda with semantic similarity for aspect-based sentiment analysis. In *2016 international joint conference on neural networks (IJCNN)*, pages 4465–4473. IEEE, 2016.
- [87] Mohamad Syahrul Mubarak, Adiwijaya, and Muhammad Dwi Aldhi. Aspect-based sentiment analysis to review products using naïve bayes. In *AIP Conference Proceedings*, volume 1867, page 020060. AIP Publishing, 2017.
- [88] Thien Hai Nguyen and Kiyooki Shirai. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514, 2015.
- [89] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. A hierarchical model of reviews for aspect-based sentiment analysis. *arXiv preprint arXiv:1609.02745*, 2016.
- [90] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, 2016.
- [91] Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy. Deepstance at semeval-2016 task 6: Detecting stance in tweets using character and word-level cnns. *arXiv preprint arXiv:1606.05694*, 2016.

- [92] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. Topical stance detection for twitter: A two-phase lstm model using attention. In *European Conference on Information Retrieval*, pages 529–536. Springer, 2018.
- [93] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. Stance classification with target-specific neural attention networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, 2017.
- [94] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 63–72. ACM, 2012.
- [95] David Vilares and Yulan He. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1582, 2017.
- [96] Sorin A Matei and Kerk F Kee. Computational communication research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1304, 2019.
- [97] Matthew Brook O’Donnell and Emily B Falk. Big data under the microscope and brains in social context: Integrating methods from computational social science and neuroscience. *The Annals of the American Academy of Political and Social Science*, 659(1):274–289, 2015.
- [98] William Stephenson. The study of behavior; q-technique and its methodology. *Psychology*, 2(9), 1953.
- [99] Carolyn B Mervis and Eleanor Rosch. Categorization of natural objects. *Annual review of psychology*, 32(1):89–115, 1981.
- [100] Rick A Adams, Quentin JM Huys, and Jonathan P Roiser. Computational psychiatry: towards a mathematically informed understanding of mental illness. *J Neurol Neurosurg Psychiatry*, 87(1):53–63, 2016.
- [101] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [102] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto. From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. *Image and Vision Computing*, 65:15–22, 2017.
- [103] Lai-Kuan Wong and Kok-Lim Low. Saliency-enhanced image aesthetics class prediction. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 997–1000. IEEE, 2009.
- [104] Xin Jin, Jingying Chi, Siwei Peng, Yulu Tian, Chaochen Ye, and Xiaodong Li. Deep image aesthetics classification using inception modules and fine-tuning connected layer. In *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*, pages 1–6. IEEE, 2016.
- [105] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876. ACM, 2014.

- [106] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 907–910. ACM, 2015.
- [107] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 159–168. ACM, 2015.
- [108] Delia Fernandez, Alejandro Woodward, Victor Campos, Xavier Giró-i Nieto, Brendan Jou, and Shih-Fu Chang. More cat than cute?: interpretable prediction of adjective-noun pairs. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, pages 61–69. ACM, 2017.
- [109] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151, 2011.
- [110] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [111] Rémi Lebret, Pedro O Pinheiro, and Ronan Collobert. Phrase-based image captioning. *arXiv preprint arXiv:1502.03671*, 2015.
- [112] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.
- [113] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics, 2011.
- [114] K. Ramnath, S. Baker, L. Vanderwende, M. El-Saban, S. N. Sinha, A. Kannan, N. Hassan, M. Galley, Y. Yang, D. Ramanan, A. Bergamo, and L. Torresani. Autocaption: Automatic caption generation for personal photos. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1050–1057, March 2014.
- [115] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. Rich image captioning in the wild. *CoRR*, abs/1603.09016, 2016.
- [116] Alex Mathews, Lexing Xie, and Xuming He. SentiCap: generating image descriptions with sentiments. In *The Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, Phoenix, USA, feb 2016.
- [117] City of Chicago. *Strategic Subject List*, 2017. <https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np>.
- [118] Karen Sheley. *Statement on Predictive Policing in Chicago*. ACLU of Illinois, Jun 2017. <http://www.aclu-il.org/en/press-releases/statement-predictive-policing-chicago>.

- [119] Christine Schmidt. *Holding algorithms (and the people behind them) accountable is still tricky, but doable*. Nieman Lab, Mar 2018. <http://nie.mn/2ucDw2>.
- [120] Matthew S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115 – 125, 2014.