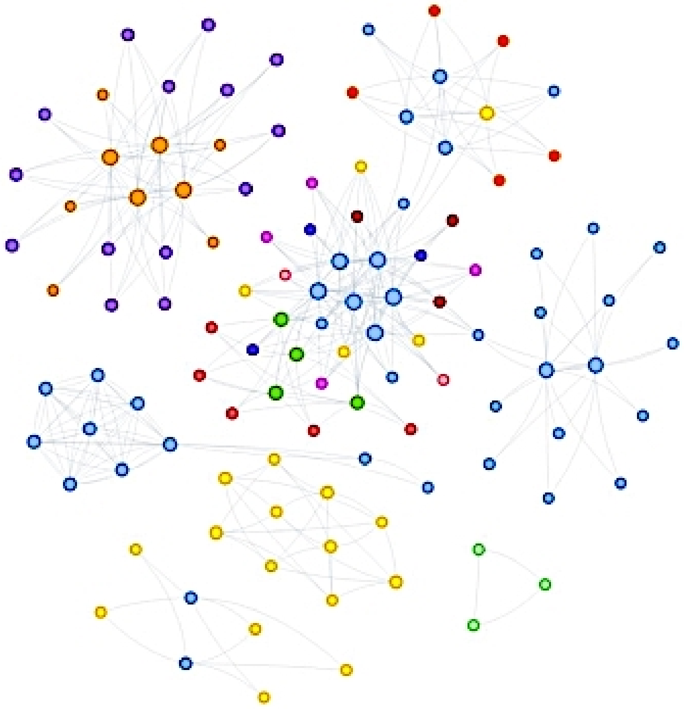
DISSERTATION

FROM SCIENTIFIC PUBLICATIONS TO COMMUNITY INSIGHTS

A Comprehensive Framework for Analyzing Academic Communities using Scientific Publications



Thesis approved by the
Department of Computer Science
University of Kaiserslautern-Landau
for the award of the Doctoral Degree

DOCTOR OF ENGINEERING (DR.-ING.)

to

Syed Tahseen Raza Rizvi

Date of the viva: 01.02.2023
Dean: Prof. Dr. Christoph Garth

Reviewers:
Prof. Dr. Prof. h.c. Andreas Dengel
Prof. Dr. Sebastian Vollmer

Rheinland-Pfälzische
Technische Universität
Kaiserslautern
Landau

DE-386

Syed Tahseen Raza Rizvi: *From Scientific Publications to Community Insights* – A Comprehensive Framework for Analyzing Academic Communities using Scientific Publications

SUPERVISOR:
Prof. Dr. Prof. h.c. Andreas Dengel

CONTACT INFORMATION:
syed_tahseen_raza.rizvi@dfki.de

Do not let your difficulties fill you with anxiety, after all it is only in the darkest nights that stars shine more brightly.

— Ali Ibne-Abi-Talib

To my loving Parents

ABSTRACT

Scientific research plays a crucial role in the development of a society. With ever-increasing volumes of scientific publications are now making it extremely challenging to analyze and maintain insights into the scientific communities like collaboration or citation trends and evolution of interests etc. This thesis is an effort towards using scientific publications to provide detailed insights into a scientific community from a range of aspects. The contribution of this thesis is five-fold.

Firstly, this thesis proposes approaches for automatic information extraction from scientific publications. The proposed layout-based approach for this purpose is inspired by how human beings perceive individual references relying only on visual queues. The proposed approach significantly outperforms the existing text-based techniques and is independent of any domain or language.

Secondly, this thesis tackles the problem of identifying meaningful topics from a given publication as the keywords provided in the publication are not always accurate representatives of the publication topic. To rectify this problem, this thesis proposes a state-of-the-art keywords extraction approach that employs a domain ontology along with the detected keywords to perform topic modeling for a given set of publications.

Thirdly, this thesis analyses the disposition of each citation to understand its true essence. For this purpose, we proposes a transformer-based approach for analyzing the impact of each citation appearing in a scientific publication. The impact of a citation can be determined by the inherent sentiment and intent of a citation, which refers to the assessment and motive of an author towards citing a scientific publication.

Furthermore, this thesis quantifies the influence of a research contributor in a scientific community by introducing a new semantic index for researchers that takes both quantitative and qualitative aspects of a citation into account to better represent the prestige of a researcher in a scientific community. Semantic Index is also evaluated for conformity to the guidelines and recommendations of various research funding organizations to assess the impact of a researcher.

In this thesis, all of the aforementioned aspects are packaged together in a single framework called Academic Community Explorer (ACE) 2.0, which automatically extracts and analyzes information from scientific publications and visualizes the insights using several interactive visualizations. These visualizations provide an instant glimpse into the scientific communities from a wide range of aspects with different granularity levels.

ACKNOWLEDGMENTS

The time during my Ph.D. has provided me with various learning opportunities and unique experiences in life. I learned to plan my work independently and in collaboration with my department colleagues. It was indeed a life-changing experience.

I am very grateful to Prof. Dr. Andreas Dengel for investing their confidence in me and providing this opportunity to complete my Ph.D. under their supervision. I have always admired his vision and interest in my thesis topic. My Ph.D. would never be possible without his support. His continuous guidance always led me through difficult times. I would also like to thank Dr. Sheraz Ahmed for our discussions on various topics of my Ph.D. and for pushing me to finish my thesis in time. A special thanks to Prof. Dr. Sebastian Vollmer for investing their time and providing valuable feedback to improve the quality of my dissertation. Furthermore, I am also thankful to Prof. Dr. Klaus Schneider for heading the proceedings of the Ph.D. committee.

I am very grateful to all my colleagues in the SDS department for their great discussions on interesting ideas that resulted in excellent collaborations. I highly appreciate the collaborations we had and look forward to more collaborations in the future. I feel lucky to have great friends in professional and personal circles who always supported me every step of the way. I like to extend my special thanks to Dr. Syed Ali Abbas Jaffari for always keeping me motivated, even through stressful times.

I am thankful to my mentors, Prof. Seiichi Uchida (Kyushu University, Japan) and Steffen Erkel (Bosch Germany), for providing me with the opportunity for a research stay and internship. I learned a lot about interesting research problems and experienced cultural diversity during these research stays. These were indeed very memorable and lifelong experiences. Furthermore, I would like to thank all my teachers for recognizing my potential and guiding me in the right direction. These crucial decisions had a butterfly effect on my life as it unfolded. This day is only possible because of their attention to detail and exceptional hard work.

The progress I made would not have been possible without prayers, sacrifices, and immense support from my family. I will always be grateful to my parents for investing their trust in me and enabling me to work towards realizing the vision of our grandfather (Dada Abbu). I am fortunate to have the best brothers in the world, who always brighten up my day, even with a few pleasant words. Last but not least, I am very thankful to my wife for her patience and unconditional support throughout my Ph.D. Thank you for always

being there to encourage me and motivating me every step of the way. A special thanks to my son, who warms my heart and fills me with energy every morning.

— Thank you, Tahseen

PUBLICATIONS AS PART OF THIS THESIS

Parts of the research and material (including figures, tables and algorithms) in this thesis have already been published in:

S. T. R. Rizvi, A. Dengel, and S. Ahmed. "A Hybrid Approach and Unified Framework for Bibliographic Reference Extraction." In: *IEEE Access* 8 (2020). doi: 10.1109/ACCESS.2020.3042455.

S. T. R. Rizvi, A. Lucieri, A. Dengel, and S. Ahmed. "Benchmarking Object Detection Networks for Image Based Reference Detection in Document Images." In: *2019 Digital Image Computing: Techniques and Applications, DICTA 2019* (2019). doi: 10.1109/DICTA47822.2019.8945991.

A. Lauscher, K. Eckert, L. Galke, A. Scherp, S. T. R. Rizvi, S. Ahmed, A. Dengel, P. Zumstein, and A. Klein. "Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph." In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (2018). doi: 10.1145/3197026.3197050.

M. Beck*, S. T. R. Rizvi*, A. Dengel, and S. Ahmed. "From automatic keyword detection to ontology-based topic modeling." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12116 LNCS (2020). doi: 10.1007/978-3-030-57058-3_32.

D. Mercier*, S. T. R. Rizvi*, V. Rajashekar, A. Dengel, and S. Ahmed. "ImpactCite: An XLNet-based solution enabling qualitative citation impact analysis utilizing sentiment and intent." In: *ICAART 2021 - Proceedings of the 13th International Conference on Agents and Artificial Intelligence* 2 (2021), pp. 159–168. doi: 10.5220/0010235201590168.

D. Mercier*, S. T. R. Rizvi*, V. Rajashekar*, S. Ahmed, and A. Dengel. "Utilizing Out-Domain Datasets to Enhance Multi-Task Citation Analysis." In: *Agents and Artificial Intelligence*. Cham: Springer International Publishing, 2022, pp. 113–134.

S. T. R. Rizvi, S. Ahmed, and A. Dengel. "A Comprehensive tool for Automatic Extraction, Analysis, and Digital Profiling of the Entities in Scientific Communities." In: *Social Network Analysis and Mining* (2023). Currently under Review.

PUBLICATIONS NOT PART OF THIS THESIS

A. Lucieri*, H. Sabir*, S. A. Siddiqui*, S. T. R. Rizvi*, B. K. Iwana, S. Uchida, A. Dengel, and S. Ahmed. "Benchmarking Deep Learning Models for Classification of Book Covers." In: *SN Computer Science* 1 (3 May 2020). DOI: 10.1007/s42979-020-00132-z.

J. Younas, S. T. R. Rizvi, M. Malik, F. Shafait, P. Lukowicz, and S. Ahmed. "FFD: Figure and Formula Detection from Document Images." In: *2019 Digital Image Computing: Techniques and Applications, DICTA 2019* (2019). DOI: 10.1109/DICTA47822.2019.8945972.

S. Siddiqui, I. Fateh, S. T. R. Rizvi, A. Dengel, and S. Ahmed. "DeepTabStR: Deep learning based table structure recognition." In: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (2019). DOI: 10.1109/ICDAR.2019.00226.

S. T. R. Rizvi, D. Mercier, S. Agne, S. Erkel, A. Dengel, and S. Ahmed. "Ontology-based Information Extraction from Technical Documents." In: *Proceedings of the 10th International Conference on Agents and Artificial Intelligence* (2018). DOI: 10.5220/0006596604930500.

CONTENTS

III DATA ANALYSIS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACE	Academic Community Explorer
OCR	Optical Character Recognition
CAG	Connectivity Aware Graph
CoCoNoW	Collective Connectivity-Aware Node Weight
CSO	Computer Science Ontolog
MOF	Minimal Occurrence Frequency
AOF	Average Occurrence Frequency
KECNW	Keyword Extraction using Collective Node Weight
NER	Node and Edge Rank
ICDAR	International Conference on Document Analysis and Recognition
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
BERT	Bidirectional Encoder Representations from Transformers
ALBERT	A Lite Bidirectional Encoder Representations from Transformers
GloVe	Global Vectors
NLP	Natural Language Processing
SVM	Support Vector Machine
CSC	Citation Sentiment Corpus
DORA	Declaration on Research Assessment
PDF	Portable Document Format

Part I

INTRODUCTION

1 INTRODUCTION

1.1 MOTIVATION

Ever since the dawn of the human race, humanity has constantly strived to make progress in every aspect of life. Nowadays, the progress of a society is generally measured in terms of research. Scientific research plays a vital role in the progress of a society. People from such communities publish their scientific contributions in formal publications. These scientific publications are widely published at scientific venues such as conferences, workshops, or journals that focus on a collection of pertinent topics. Each scientific publication focuses on solving a challenging problem with its unique approach. At times, a scientific problem was proved to be profoundly demanding and required a joint effort from multiple publications making modest enhancements. Therefore, ultimately building on the scientific contributions of their fellow researchers.

For the ventures that include extending a previous research work, the upcoming publications quote the existing literature which undertook the endeavor to solve the task at hand. It is performed to give credit to the prior researchers for their achievements and contributions to the scientific community. Such referencing of the published literature is known as a citation. Citations play a vital role to quantify the significance and impact of the scientific contributions by researchers. Citations ensure that the relevant researcher is getting proper credit for their research irrespective of the fact that it was either a significant or minor contribution.

For a long time, the citations were considered as a mere quantitative measure. However, some initial efforts [58] brought attention to the context of each citation which could potentially differentiate one citation from the other. In addition to the existing aspect of giving credit to prior research work, the citations now had an additional aspect where citations serve as a system to validate the integrity of research work. For instance, a publication improves the results or points out the limitations of an existing approach. In principle, both scenarios result in a citation however, their respective contexts are significantly different. Hence, citations can also be used as a tool to highlight the shortcomings of a scientific contribution.

With the ever-increasing number of publications, it became more challenging to evaluate the impact of a researcher. For this purpose, an author index [54] was introduced to quantify the impact of research performed by an individual in a scientific community. It immediately became popular and was widely adopted by evaluators. Over time, the limitations of the H-index got identified. Later, several author indexes were proposed to mitigate the shortcomings of previous indexes. However, the subsequent indexes had their inherent limitations deeming them either not robust or suitable for all domains in science.

In recent times, citations have had a decisive role in financing research work. As several activities in research like funding, scholarships, and recruitment decisions highly rely on the impact of the research performed by individuals, research groups, or organizations. Therefore, it is more important than ever to assess the impact of research performed by individuals. With the increasing volume of publications, it became infeasible to manually analyze all the scientific publications as Peer-review to access the quality is a slow and tedious process. To circumnavigate this problem, the community proposed several tools [48, 110] to analyze scientific publications. However, these analysis tools provided very shallow insights into the scientific publications and were highly dependent on outdated author indexes that introduce their intrinsic limitations.

This thesis proposes a comprehensive framework called Academic Community Explorer (ACE) for analyzing scientific publications. It comprises modules that automatically detect and extract publication metadata along with all the interesting data like bibliographic references, keywords, etc. The second module performs a detailed analysis of all the extracted information to identify different trends and patterns. The analysis is performed from multiple aspects, ranging from an individual researcher to a community level. This thesis introduces several author roles to identify influential researchers from different perspectives. This research work also includes the study to predict the citation sentiment and intent. Furthermore, this thesis also mitigates the limitations of existing author indexes by proposing a novel semantic index that covers more meaningful aspects of a researcher's work in a scientific community. The final block displays all the extracted data and trends using an interactive visualization engine that provides instant, detailed insights into the research performed by an individual or a scientific community.

1.2 RESEARCH QUESTIONS & OBJECTIVES

The research questions aimed in this thesis and their respective objectives are described as follows:

1. **Can we exploit layout features to extract bibliographic references with better performance than traditional text-based reference extraction approaches?**

Objective: Explore several object detection architectures to detect bibliographic references using layout features. Compare the performance of reference detection models to select the best-performing model for the proposed pipeline. Additionally, perform a comparative evaluation of the layout detection approach with the baseline state-of-the-art text-based approaches.

2. **How can we identify the nature of individual community interactions in a community?**

   Objective: Propose a method that helps to analyze the quality and purpose of community interaction i.e., citations. The quality and purpose of a given citation refer to the sentiment and intent, respectively. Additionally, compare the performance of the proposed models to existing sentiment and intent classification models.

3. **What aspects can be integrated into the author's index in addition to quantitative aspects like citation count?**

   Objective: The first objective is to investigate the limitations of the widely used existing author indexes. Secondly, propose a novel author semantic index that mitigates the limitations and disadvantages of the previous indexes. Furthermore, evaluate the proposed author semantic index using the guidelines and principles defined by the scientific community standards.

4. **Is it possible to identify Influential individuals in a community using scientific community data?**

   Objective: Propose a comprehensive pipeline that extracts the detected citations from a given publication and analyzes it to determine the citing publication and the cited target publications. Furthermore, use the extracted metadata from both publications to identify the collaborating and citing authors. Eventually, that approach uses all extracted data to analyze each author from different community roles. Additionally, propose meaningful community roles and identify leaders for each proposed role in a scientific community.

5. **Is it possible to identify cliques in the community using only publications data?**

   Objective: Propose a comprehensive tool to extract metadata from scientific publications and analyze it to identify the entities in the data and the relation among those entities. Later, identify any trends in the interaction of the entities. Additionally, visualize all the extracted entities and their interactions to identify a tightly knitted subgroup in the scientific community.

1.3 CONTRIBUTIONS

The contributions of this thesis are as follows:

1. This thesis proposes a deep learning-based Bibliographic Reference Detection approach called DeepBiRD. Unlike existing state-of-the-art text-based methods, the proposed method relies on the layout features to identify the references from a scanned scientific publication. DeepBiRD pre-processes the input image to highlight the layout features like space between characters, words, and lines to facilitate the detection mechanism. The layout features enable the DeepBiRD to be more robust and generalized, therefore, making it independent of the referencing styles, textual content, and even the language of a given scientific publication. Additionally, a dataset of 2401 scanned images from scientific publications containing 38863 annotated references was also publicly released along with this work. The proposed approach outperformed both the text and image-based reference detection methods by a significant margin. DeepBiRD was evaluated on two different datasets and achieved a mean average precision of 97.59% and 98.56% at Intersection over Union (IoU) of 50. The detailed evaluations of the proposed approach validate the first hypothesis that replacing textual features with layout features to identify bibliographic references indeed provides a performance boost and is, therefore, more effective.

2. For the task of Keywords and Topic detection from scientific publications, this study proposes a two-stage approach. The first stage of the proposed method is called Collective Connectivity aware Node Weight (CoCoNoW). It is responsible for identifying keywords from a given scientific text corpus. CoCoNoW builds a connectivity-aware graph from the input publication where each node represents a potential keyword. Initially, each node in the connectivity-aware graph is assigned a weight. Later, the nodes are sorted based on their node weight in the order of relevance. The second stage of the proposed approach specializes in identifying the topic of a given publication. For this purpose, this study employs a domain ontology that encompasses the complete taxonomy of that domain. The use of ontology enables this study to be independent of any domain, as replacing the ontology makes this approach adaptable to any provided field of study. CoCoNoW was evaluated extensively on three publicly available datasets, where it consistently outperformed all other state-of-the-art approaches on all three datasets.

3. The impact of a scientific publication can be well understood using the citations received by that publication. This study pro-

poses an approach called ImpactCite, which analyzes the qualitative aspects of individual citations. These aspects include the sentiment and intent of the citing publication in the literature. ImpactCite enables us to quantify the qualitative facets of citing existing literature. Which ultimately helps us to understand the overall impact of a scientific publication. There are three contributions to this study. Firstly, it benchmarked the performance of the renowned architectures for both tasks. Secondly, ImpactCite achieved state-of-the-art results by outperforming the existing approaches with an F1 score of 3.44% and 1.33% for the tasks of citation sentiment and intent sentiment classification, respectively. The third contribution of this work involves releasing a cleaned and reliable dataset for the citation sentiment analysis. This study validates the second hypothesis that it is possible to identify the nature of an individual interaction in a scientific publication using the context of the citations data.

4. In scientific communities, the quantification of performance and influence of a researcher is generally performed using the number of citations received by that researcher throughout their career. For this purpose, several renowned author citation indexes have been widely in-use nowadays. These indexes effectively cover the quantitative aspect of the citations. However, they severely lack the qualitative facet of the citations received by a fellow researcher in the research community. This study proposed a novel Semantic index for researchers which considers both quantitative and qualitative aspects of a citation. These aspects help in estimating the influence of a researcher in a scientific community. Another contribution of this study was to introduce several roles like collaboration, experts, community connection, etc. that represent certain aspects of a scientific community and serve as its pillars. Each researcher is analyzed from all these different aspects. And their comparative assessment is performed for the individual role in the community. This study validates the third and fourth hypotheses that state that considering both quantitative and qualitative aspects enable us to get a more meaningful picture and helps in better understanding the influence of a researcher for different roles in a scientific community.

5. Interactions of a scientific community are concealed in a publication due to their nature. The reason for this latent nature is because of an additional pre-processing step required to mine such interactions out of a scientific paper. This study contributed by providing a comprehensive tool called Academic Community Explorer (ACE) which is responsible for autonomously detecting and extracting relevant entities and their respective inter-

actions in the scientific community. ACE analyzes the extracted entities and their interactions to identify citation trends. Additionally, ACE computes different useful statistics for a quick insight into a scientific community. ACE is equipped with an interactive visualization engine that effectively visualizes the trends and extracted data. This work validates our last hypothesis that it is indeed feasible to identify cliques in a given scientific community.

1.4 OUTLINE

The structure of this thesis is as follows:

Chapter 1 discussed the Motivation of this work (Section 1.1) followed by Research Questions and Objectives (Section 1.2) and Contributions (Section 1.3).

The rest of the thesis is divided into three parts each of which represents a pillar of ACE 2.0 i.e., Data Extraction, Data Analysis, and Data Visualization. The data Extraction part includes two chapters. Chapter 2 performs the benchmarking of object detection models for identifying references from scientific publications. Later, that chapter introduces a novel approach for layout-based bibliographic reference detection. Chapter 3 proposes a graph-based approach for extracting keywords from scientific publications followed by using the extracted keywords along with a domain ontology to perform topic modeling.

The Data Analysis part consists of two chapters. Chapter 4 proposes a citation impact analysis approach called ImpactCite which performs sentiment and intent analysis of citations in a scientific publication. Chapter 5 discusses the contribution of this thesis towards introducing a novel qualitative author citation index that assesses the impact of research performed by an individual in a scientific community. The Data Visualization part sheds light on the interactive visualization engine of ACE. All the visualizations and their details are discussed in Chapter 6.

Chapter 7 discusses the associated research performed during this thesis. That includes three studies on different tasks like Book cover classification, Figure, and Formula detection, and Table Structure Recognition in scientific publications. Finally, Chapter 8 summarizes the conclusions of this thesis followed by the discussion of the potential future directions to extend this research work.

Part II

DATA EXTRACTION

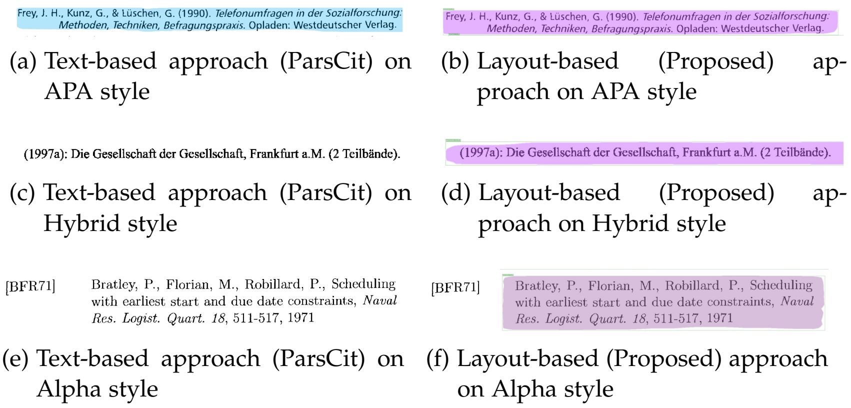
2 BIBLIOGRAPHIC REFERENCE EXTRACTION

Publications are an integral part of a scientific community. There has been a rapid progress in every field of research since the start of the 21st century, subsequently resulting in an exponential increase in the volume of scientific literature [60, 122]. Bibliographic references play a vital role for every publication as they establish the authenticity and influence of a research artifact in the scientific community.

Bibliographic references are of particular interest to library communities [68]. They play a key role in compiling library catalogs. These catalogs contain information regarding all bibliographic artifacts like books, journals, conference proceedings, magazines, and other media present in a library. For such a scale, it is not feasible to manually find and index a huge volume of references.

Resource Discovery Systems pose to be a viable solution for the libraries to further expanding their horizon by providing the indexed bibliographic data available from external resources [68]. Some resources are commercial and are thus paid by the user to take advantage of their collected data. Such platforms include Web of Science, Scopus, etc. However, even the paid subscriptions of such platforms do not ensure the access of user to the complete data from a wide range of literature published in different domains and written in several languages. According to a scientometric study, [87], both Scopus and Web of Science have mostly coverage of English journal articles from the Biomedical and Social Science domains and thus have low overall coverage for journals and articles from other domains and languages. Therefore deeming the Resource Discovery Systems a suboptimal solution for bibliographic cataloging.

The majority of the literature on the task of bibliographic reference detection consists of text-based solutions which make use of textual features like author names, publication titles, etc. in a document to detect references. Text-based approaches use a set of carefully crafted heuristics and regular expressions [29, 109] based on the position of the constituents of a reference string i.e. author names, affiliations, publisher, journal/book/conference name, year of publishing, etc. The defined heuristics and regular expressions are very sensitive to the textual features as they do not anticipate any number or special character like brackets etc. at the start of a reference string thus any such references are prone to be missed altogether. Text-based approaches are also very sensitive to the layout of the references for example if all lines of reference strings start from the same point then text-based approaches find it very hard to identify the starting and



Frey, J. H., Kunz, G., & Lüschen, G. (1990). *Telefonumfragen in der Sozialforschung: Methoden, Techniken, Befragungspraxis.* Opladen: Westdeutscher Verlag.

(a) Text-based approach (ParsCit) on APA style

Frey, J. H., Kunz, G., & Lüschen, G. (1990). *Telefonumfragen in der Sozialforschung: Methoden, Techniken, Befragungspraxis.* Opladen: Westdeutscher Verlag.

(b) Layout-based (Proposed) approach on APA style

(1997a): Die Gesellschaft der Gesellschaft, Frankfurt a.M. (2 Teilbände).

(c) Text-based approach (ParsCit) on Hybrid style

(1997a): Die Gesellschaft der Gesellschaft, Frankfurt a.M. (2 Teilbände).

(d) Layout-based (Proposed) approach on Hybrid style

[BFR71] Bratley, P., Florian, M., Robillard, P., Scheduling with earliest start and due date constraints, *Naval Res. Logist. Quart. 18*, 511-517, 1971

(e) Text-based approach (ParsCit) on Alpha style

[BFR71] Bratley, P., Florian, M., Robillard, P., Scheduling with earliest start and due date constraints, *Naval Res. Logist. Quart. 18*, 511-517, 1971

(f) Layout-based (Proposed) approach on Alpha style

Figure 2.1: Detection results of text-based and layout-based methods

ending boundaries of a reference string thus either detecting multiple references as one or identifying one reference as multiple references. With the introduction of such cases, carefully crafted heuristics become deprecated right away thus making text-based approaches less robust and eventually not generalizable. For instance, most common referencing styles like MLA and APA have author names and publication titles as the starting features of a reference string. On the other hand, there exist some rare bibliography styles in social sciences such as Alpha or hybrid Chicago in which reference strings either start with a reference identifier or publication year respectively. It is an example of a problematic case for text-based approaches because the references in the given example do not comply with the other common traditional referencing styles and are rarely used. Comparison of results from text-based and layout-based approaches on different referencing styles is shown in Fig 2.1. It can be observed that the text-based approach was unable to detect an unusual reference string as it entirely relies on textual features while overlooking other important facets i.e. layout features.

This chapter introduces an automatic, effective, and generalized approach for reference detection from document images. It works equally well for scanned and digital-born PDF documents. The proposed approach is inspired by the way how human beings perceive and identify objects. To understand this phenomenon, consider an example of an illegible blurred document containing some text. Although the document is unreadable, however, we can still identify paragraphs, bullet points, and similarly references. The underlying idea states that layout information is the key to identify different textual structures in a document even without using textual features. For this task, we employed Convolutional Neural Networks (CNN) for representation learning based on the layout of a document. This waives the dependency on textual heuristics used in the majority of the existing systems. The proposed approach is generic and is thus applicable to any bibliographic publication independent of its domain or referencing style. We also releases a benchmark dataset for bibliographic reference detection from document images. We performed benchmark tasks to evaluate the performance of *DeepBiRD* [103] from different aspects i.e. generalization, robustness, etc.

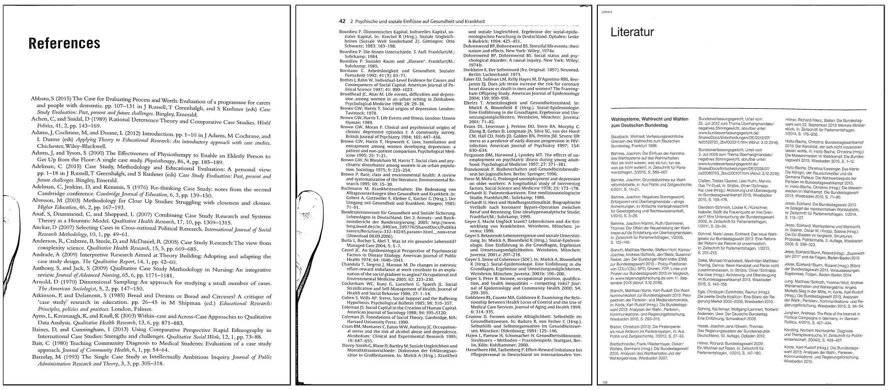
This chapter also presents a comprehensive framework called BRExSys, which encapsulates all state-of-the-art bibliographic reference detection methods under a single umbrella, allowing users to use any of the existing or proposed methods in one place. BRExSys supports scientific publications in a number of file formats i.e. born-digital PDF, Scanned PDF/images, HTML, XML etc. BRExSys provides an intuitive, user-friendly interface to facilitate the smooth processing of the input file and visualization of processed output. BRExSys is a highly customizable system as it can be tailored based on the user's requirements.

The contributions of this chapter are as follows:

- We present a novel layout driven approach for automatic reference detection from scientific publications, which effectively exploits the visual cues to firstly identify bibliographic references from a given scientific publication.
- We release a new & larger dataset for image-based reference detection which will be publicly available for the community.
- We demonstrate the superiority of the proposed approach by carrying out a series of comparative performance evaluations against the previous state-of-the-art approach.
- We present an automatic bibliographic reference detection framework called BRExSys which has integrated *DeepBiRD* [103] and other state-of-the-art text-based reference detection models to take advantage of different modalities i.e. layout and text, etc. for the task at hand.

2.1 LITERATURE REVIEW

Reference detection is a popular task among scientific community. Bibliographic reference detection is generally performed by two methods like text-based and layout-based. Most of the approaches are based on the analysis of textual content to identify references. There are several techniques employed by each text-based approach to identify references. Here we will discuss such techniques used for bibliographic reference detection, starting from the simplest and moving towards more sophisticated ones.



References

Abbato, S (2015) The Case for Evaluating Process and Worth: Evaluation of a programme for carers and people with dementia. pp. 107–131 in J Russell, T Greenhalgh, and S Kushner (eds) *Case Study Evaluation: Past, present and future challenges*. Bingley, Emerald.

Achen, C, and Snidal, D (1989) Rational Deterrence Theory and Comparative Case Studies. *World Politics*, 41, 2, pp. 143–169.

Adams, J, Cochrane, M, and Dunne, L (2012) Introduction. pp. 1–10 in J Adams, M Cochrane, and L Dunne (eds) *Applying Theory to Educational Research: An introductory approach with case studies*. Chichester, Wiley-Blackwell.

Adams, J, and Tyson, S (2000) The Effectiveness of Physiotherapy to Enable an Elderly Person to Get Up from the Floor: A single case study. *Physiotherapy*, 86, 4, pp. 185–189.

Adelman, C (2015) Case Study, Methodology and Educational Evaluation: A personal view. pp. 1–18 in J Russell, T Greenhalgh, and S Kushner (eds) *Case Study Evaluation: Past, present and future challenges*. Bingley, Emerald.

Adelman, C, Jenkins, D, and Kemmis, S (1976) Re-thinking Case Study: notes from the second Cambridge conference. *Cambridge Journal of Education*, 6, 3, pp. 139–150.

Alvesson, M (2003) Methodology for Close Up Studies: Struggling with closeness and closure. *Higher Education*, 46, 2, pp. 167–193.

Anaf, S, Drummond, C, and Sheppard, L (2007) Combining Case Study Research and Systems Theory as a Heuristic Model. *Qualitative Health Research*, 17, 10, pp. 1309–1315.

Anckar, D (2007) Selecting Cases in Cross-national Political Research. *International Journal of Social Research Methodology*, 10, 1, pp. 49–61.

Anderson, R, Crabtree, B, Steele, D, and McDaniel, R (2005) Case Study Research: The view from complexity science. *Qualitative Health Research*, 15, 5, pp. 669–685.

Andrade, A (2009) Interpretive Research Aimed at Theory Building: Adopting and adapting the case study design. *The Qualitative Report*, 14, 1, pp. 42–60.

Anthony, S, and Jack, S (2009) Qualitative Case Study Methodology in Nursing: An integrative review. *Journal of Advanced Nursing*, 65, 6, pp. 1171–1181.

Arnold, D (1970) Dimensional Sampling: An approach for studying a small number of cases. *The American Sociologist*, 5, 2, pp. 147–150.

Atkinson, P, and Delamont, S (1985) Bread and Dreams or Bread and Circuses? A critique of 'case study' research in education. pp. 26–45 in M Shipman (ed.) *Educational Research: Principles, policies and practices*. London, Falmer.

Ayres, L, Kavanaugh, K, and Knafl, K (2003) Within-case and Across-Case Approaches to Qualitative Data Analysis. *Qualitative Health Research*, 13, 6, pp. 871–883.

Baines, D, and Cunningham, I (2013) Using Comparative Perspective Rapid Ethnography in International Case Studies: Strengths and challenges. *Qualitative Social Work*, 12, 1, pp. 73–88.

Bair, C (1980) Teaching Community Diagnosis to Medical Students: Evaluation of a case study approach. *Journal of Community Health*, 6, 1, pp. 54–64.

Barzelay, M (1993) The Single Case Study as Intellectually Ambitious Inquiry. *Journal of Public Administration Research and Theory*, 3, 3, pp. 305–318.

(a) Single Column

42 2 Psychische und soziale Einflüsse auf Gesundheit und Krankheit

Bourdieu P. Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. In: Kreckel R (Hrsg.). Soziale Ungleichheiten (Soziale Welt Sonderband 2). Göttingen: Otto Schwartz; 1983: 183–198.

Bourdieu P. Die feinen Unterschiede. 3. Aufl. Frankfurt/M.: Suhrkamp; 1984.

Bourdieu P. Sozialer Raum und „Klassen". Frankfurt/M.: Suhrkamp; 1985.

Bormann C. Arbeitslosigkeit und Gesundheit. Sozialer Fortschritt 1992; 41 (3): 63–71.

Brehm J, Rahn W. Individual-Level Evidence for Causes and Consequences of Social Capital. American Journal of Political Science 1997; 41: 999–1023.

Broadhead JC, Abas M. Life events, difficulties and depression among women in an urban setting in Zimbabwe. Psychological Medicine 1998; 28: 29–38.

Brown GW, Harris T. Social origins of depression. London: Tavistock; 1978.

Brown GW, Harris T. Life Events and Illness. London: Unwin Hyman; 1989.

Brown GW, Moran P. Clinical and psychosocial origins of chronic depressive episodes I: A community survey. British Journal of Psychiatry 1994; 165: 447–456.

Brown GW, Harris T, Hepworth C. Loss, humiliation and entrapment among women developing depression: a patient and non-patient comparison. Psychological Medicine 1995; 25: 7–21.

Brown GW, Ni Bhrolchain M, Harris T. Social class and psychiatric disturbance among women in an urban population. Sociology 1975; 9: 225–254.

Brown P. Race, class and environmental health: A review and systematization of the literature. Environmental Research 1995; 69: 15–30.

Buchmann M. Krankheitsverhalten: Die Bedeutung von Alltagsvorstellungen über Gesundheit und Krankheit. In: Gebert A, Gutzwiller F, Kleiber C, Kocher G (Hrsg.). Der Umgang mit Gesundheit und Krankheit. Horgen; 1985: 71–91.

Bundesministerium für Gesundheit und Soziale Sicherung. Lebenslagen in Deutschland. Der 2. Armuts- und Reichtumsbericht der Bundesregierung; 2005; http://www.bmg.bund.de/cln_040/nn_599776/SharedDocs/Publikationen/Berichte/a-332-10245,param=.html__nnn=true (Download 08.09.2006).

Burla L, Bucher S, Abel T. Was ist ein gesunder Lebensstil? Managed Care 2004; 5: 5–7.

Cassel JC. An Epidemiological Perspective of Psychosocial Factors in Disease Etiology. American Journal of Public Health 1974; 64: 1040–1043.

Chandola T, Siegrist J, Marmot M. Do changes in extrinsic effort-reward imbalance at work conribute to an explanation of the social gradient in angina? Occupational and Environmental Medicine 2005; 62: 223–230.

Cockerham WC, Kunz G, Lueschen G, Spaeth JL. Social Stratification and Self-Management of Health. Journal of Health and Social Behavior 1986; 27: 1–14.

Cohen S, Wills AP. Stress, Social Support and the Buffering Hypothesis. Psychological Bulletin 1985; 98: 310–357.

Coleman JS. Social Capital in the Creation of Human Capital. American Journal of Sociology 1988; 94: S95–S120.

Coleman JS. Foundations of Social Theory. Cambridge, MA: Harvard University Press; 1990.

Crum RM, Muntaner C, Eaton WW, Anthony JC. Occupational stress and the risk of alcohol abuse and dependence. Alcoholism: Clinical and Experimental Research 1995; 19: 647–655.

Davey-Smith G, Blane D, Bartley M. Soziale Ungleichheit und Mortalitätsunterschiede: Diskussion der Erklärungsansätze in Großbritannien. In: Mielck A (Hrsg.). Krankheit und soziale Ungleichheit. Ergebnisse der sozial-epidemiologischen Forschung in Deutschland. Opladen: Leske & Budrich; 1994: 425–451.

Dohrenwend BP, Dohrenwend BS. Stressful life events: their nature and effects. New York: Wiley; 1974a.

Dohrenwend BP, Dohrenwend BS. Social status and psychological disorder: A causal inquiry. New York: Wiley; 1974b.

Durkheim E. Der Selbstmord (frz. Original: 1897). Neuwied, Berlin: Luchterhand; 1973.

Eaker ED, Sullivan LM, Kelly Hayes M, D'Agostino RBS, Benjamin EJ. Does job strain increase the risk for coronary heart disease or death in men and women? The Framingham Offspring Study. American Journal of Epidemiology 2004; 159: 950–958.

Elkeles T. Arbeitslosigkeit und Gesundheitszustand. In: Mielck A, Bloomfield K (Hrsg.). Sozial-Epidemiologie. Eine Einführung in die Grundlagen, Ergebnisse und Umsetzungsmöglichkeiten. Weinheim, München: Juventa; 2001: 71–82.

Evans DL, Leserman J, Perkins DO, Stern RA, Murphy C, Zheng B, Gettes D, Longmate JA, Silva SG, van der Horst CM, Hall CD, Folds JD, Golden RN, Petitto JM. Severe life stress as a predictor of early disease progression in HIV-infection. American Journal of Psychiatry 1997; 154: 630–634.

Fergusson DM, Horwood J, Lynskey MT. The effects of unemployment on psychiatric illness during young adulthood. Psychological Medicine 1997; 27: 371–381.

Franzkowiak P. Risikoverhalten und Gesundheitsbewußtsein bei Jugendlichen. Berlin: Springer; 1996.

Frese M, Mohr G. Prolonged unemployment and depression on older workers: A longitudinal study of intervening factors. Social Science and Medicine 1978; 25: 173–178.

Gerhardt U. Patientenkarrieren. Eine medizinsoziologische Studie. Frankfurt/M.: Suhrkamp; 1986.

Gerhardt U. Herz und Handlungsrationalität. Biographische Verläufe nach koronarer Bypass-Operation zwischen Beruf und Berentung. Eine idealtypenanalytische Studie. Frankfurt/M.: Suhrkamp; 1999.

Geyer S. Macht Unglück krank? Lebenskrisen und die Entwicklung von Krankheiten. Weinheim, München: Juventa; 1999.

Geyer S. Belastende Lebensereignisse und soziale Unterstützung. In: Mielck A, Bloomfield K (Hrsg.). Sozial-Epidemiologie. Eine Einführung in die Grundlagen, Ergebnisse und Umsetzungsmöglichkeiten. Weinheim, München: Juventa; 2001 a: 207–218.

Geyer S. Sense of Coherence (SOC). In: Mielck A, Bloomfield K (Hrsg.). Sozial-Epidemiologie. Eine Einführung in die Grundlagen, Ergebnisse und Umsetzungsmöglichkeiten. Weinheim, München: Juventa; 2001b: 196–206.

Geyer S, Peter R. Income, occupational position, qualification, and health inequalities – competing risks? Journal of Epidemiology and Community Health 2000; 54: 299–305.

Goldsteen RL, Counte MA, Goldsteen K. Examining the Relationship Between Health Locus of Control and the Use of Medical Care Services. Journal of Aging and Health 1994; 6: 314–335.

Grunow D. Formen sozialer Alltäglichkeit: Selbsthilfe im Gesundheitswesen. In: Badura B, von Ferber C (Hrsg.). Selbsthilfe und Selbstorganisation im Gesundheitswesen. München: Oldenbourg; 1981: 125–146.

Hajen L, Paetow H, Schumacher H. Gesundheitsökonomie. Strukturen – Methoden – Praxisbeispiele. Stuttgart, Berlin, Köln: Kohlhammer; 2000.

Hasselhorn HM, Tackenberg P. Effort-Reward Imbalance bei Pflegepersonal in Deutschland im internationalen Ver-

(b) Double Column

Literatur

Literatur

Wahlsysteme, Wahlrecht und Wahlen zum Deutschen Bundestag

Bausback, Winfried: Verfassungsrechtliche Grenzen des Wahlrechts zum Deutschen Bundestag, Frankfurt 1998.

Behnke, Joachim: Der Einfluss der Kenntnis des Wahlsystems auf das Wahlverhalten: Weil sie nicht wissen, was sie tun, tun sie, was sie nicht wollen?, in: Zeitschrift für Parlamentsfragen, 3/2015, S. 588–607.

Behnke, Joachim: Grundsätzliches zur Wahlreformdebatte, in: Aus Politik und Zeitgeschichte, 4/2011, S. 14–21.

Behnke, Joachim: Negatives Stimmgewicht, Erfolgswert und Überhangmandate – einige Anmerkungen, in: Kritische Vierteljahresschrift für Gesetzgebung und Rechtswissenschaft, 1/2010, S. 3–28.

Behnke, Joachim/Kamm, Ruth/Sommerer, Thomas: Der Effekt der Neueinteilung der Wahlkreise auf die Entstehung von Überhangmandaten, in: Zeitschrift für Parlamentsfragen, 1/2003, S. 122–145.

Bianchi, Matthias/Bender, Steffen/Hohl, Karina/Jüschke, Andreas/Schoofs, Jan/Steitz, Susanne/Treibel, Jan: Der Duisburger-Wahl-Index (DWI) zur Bundestagswahl 2013 – Policy-Positionen von CDU/CSU, SPD, Grünen, FDP, Linke und Piraten zur Bundestagswahl 2013 im Vergleich, in: www.regierungsforschung.de vom 11. September 2013 (Abruf: 3.12.2016).

Bianchi, Matthias/Korte, Karl-Rudolf: Die Wahlkommunikation zur Bundestagswahl 2013: Perspektiven der Parteien- und Mediendemokratie, in: Korte, Karl-Rudolf (Hrsg.): Die Bundestagswahl 2013: Analysen der Wahl-, Parteien-, Kommunikations- und Regierungsforschung, Wiesbaden 2015, S. 293–315.

Bieber, Christoph (2012): Die Piratenpartei als neue Akteurin im Parteiensystem, in: Aus Politik und Zeitgeschichte, 7/2012, S. 27–33.

Brettschneider, Frank/Niedermayer, Oskar/Weßels, Bernhard (Hrsg.): Die Bundestagswahl 2005. Analysen des Wahlkampfes und der Wahlergebnisse, Wiesbaden 2007.

Bundesverfassungsgericht, Urteil vom 25. Juli 2012 zum Thema Überhangmandate/negatives Stimmgewicht, abrufbar unter: www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2012/07/fs20120725_2bvf000311.html (Abruf: 3.12.2016).

Bundesverfassungsgericht, Urteil vom 3. Juli 2008 zum Thema Überhangmandate/negatives Stimmgewicht, abrufbar unter: www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/2008/07/cs20080703_2bvc000107.html (Abruf: 3.12.2016).

Claßen, Tobias/Guckel, Lisa/Kuhn, Marvin: Das TV-Duell, in: Strijbis, Oliver/Schnapp, Kai-Uwe (Hrsg.): Aktivierung und Überzeugung im Bundestagswahlkampf 2013, Wiesbaden 2015, S. 159–176.

Davidson-Schmich, Louise K./Kürschner, Isabelle: Stößt die Frauenquote an ihre Grenzen? Eine Untersuchung der Bundestagswahl 2009, in: Zeitschrift für Parlamentsfragen, 1/2011, S. 25–34.

Dehmel, Niels/Jesse, Eckhard: Das neue Wahlgesetz zur Bundestagswahl 2013. Eine Reform der Reform der Reform ist unvermeidlich, in: Zeitschrift für Parlamentsfragen, 1/2013, S. 201–213.

Dietel, Michael/Knackstedt, Maximilian Matthies/Thering, Dennis: Wenn Kandidat und Partei nicht zusammenpassen, in: Strijbis, Oliver/Schnapp, Kai-Uwe (Hrsg.): Aktivierung und Überzeugung im Bundestagswahlkampf 2013, Wiesbaden 2015, S. 143–158.

Egle, Christoph/Zohlnhöfer, Reimut (Hrsg.): Die zweite Große Koalition. Eine Bilanz der Regierung Merkel 2005–2009, Wiesbaden 2010.

Gehring, Kai/Ismayr, Wolfgang/Lammert, Norbert/Andersen, Uwe: Der Deutsche Bundestag. Eine Einführung, Schwalbach 2008.

Hesse, Joachim Jens/Ellwein, Thomas: Das Regierungssystem der Bundesrepublik Deutschland, 10. Auflage, Opladen 2012.

Hilmer, Richard: Bundestagswahl 2009: Ein Wechsel auf Raten, in: Zeitschrift für Parlamentsfragen, 1/2010, S. 147–180.

Hilmer, Richard/Merz, Stefan: Die Bundestagwahl vom 22. September 2013: Merkels Meisterstück, in: Zeitschrift für Parlamentsfragen, 1/2014, S. 175–206.

Holtz-Bacha, Christina: Bundestagswahlkampf 2013: Der Kandidat, der sich nicht inszenieren lassen wollte, in: Holtz-Bacha, Christina (Hrsg.): Die Massenmedien im Wahlkampf. Die Bundestagswahl 2013, Wiesbaden 2015, S. 1–12.

Holtz-Bacha, Christina/Lessinger, Eva-Maria: Die Königin, der Rausschmeißer und die Gemeine Fitzlaus: Die Wahlwerbespots der Parteien im Bundestagswahlkampf 2013, in: Holtz-Bacha, Christina (Hrsg.): Die Massenmedien im Wahlkampf. Die Bundestagswahl 2013, Wiesbaden 2015, S. 71–90.

Jesse, Eckhard: Die Bundestagswahl 2013 im Spiegel der repräsentativen Wahlstatistik, in: Zeitschrift für Parlamentsfragen, 1/2014, S. 113–127.

Jesse, Eckhard: Wahlsysteme und Wahlrecht, in: Gabriel, Oscar W./Kropp, Sabine (Hrsg.): Die EU-Staaten im Vergleich. Strukturen, Prozesse, Politikinhalte, 3. Auflage, Wiesbaden 2008, S. 299–323.

Jesse, Eckhard/Sturm, Roland (Hrsg.): „Superwahljahr 2011" und die Folgen, Baden-Baden 2012.

Jesse, Eckhard/Sturm, Roland (Hrsg.): Bilanz der Bundestagswahl 2013. Voraussetzungen, Ergebnisse, Folgen, Baden-Baden 2014.

Jung, Matthias/Schroth, Yvonne/Wolf, Andrea: Wählerverhalten und Wahlergebnis: Angela Merkels Sieg in der Mitte, in: Korte, Karl-Rudolf (Hrsg.), Die Bundestagswahl 2013. Analysen der Wahl-, Parteien-, Kommunikations- und Regierungsforschung, Wiesbaden 2015, S. 34–54.

Jungherr, Andreas: The Role of the Internet in Political Campaigns in Germany, in: German Politics, 4/2015, S. 427–434.

Kersting, Norbert: Nichtwähler. Diagnose und Therapieversuche, in: Zeitschrift für Politikwissenschaft, 2004/3, S. 403–427.

Korte, Karl-Rudolf (Hrsg.): Die Bundestagswahl 2013: Analysen der Wahl-, Parteien-, Kommunikations- und Regierungsforschung, Wiesbaden 2015.

156

(c) Triple Column

Figure 2.2: Samples of different layouts from input files

2.1.1 *Text-based Approaches*

The simplest of the text-based reference detection techniques employ regular expressions and carefully crafted heuristics [29] for this task. Such approaches are mostly not considered as an optimal solution because of their limited coverage. For example, *MLA* and *APA* are the most common referencing styles in which a reference string starts with author names. To detect such references, adopted heuristics will look for comma-separated author names at the start of the reference string. The drawback of such an approach is that it will be unable to detect a reference if it does not comply with the defined heuristics i.e. reference string with *Alpha* style where reference string starts with a custom ID. Every domain has its unique referencing style and sometimes there are multiple referencing styles within one domain. Such challenges make simple approaches unsuitable for this complex task.

Citation-Parser [29] is a typical example of heuristics based tool. To identify components of bibliographic reference string i.e. authors, title, conference/journal, etc., it employs a set of carefully designed heuristics. Sautter *et al.* [109] proposed a tool named *RefParse* which exploits similarities between individual reference strings to identify different referencing style for parsing a reference string. Perl also provided an extension named *Biblio* [16] for parsing and extracting reference string metadata. Chen *et al.* [28] proposed *BibPro*, an approach that identifies citation style by matching it with referencing styles available in its database and then uses gene sequence alignment technique to identify components of reference strings. *AnyStyle-Parser* [6] is another example of a tool which identifies bibliographic references using heuristics. *PDFSSA4MET* [98] proposed a slightly different approach to identify references in a Born-digital PDF. In this approach, textual PDF is firstly converted into an XML file. Then by employing pattern matching mechanisms, syntactic and structural analysis of XML is performed to identify the reference section.

Ahmed *et al.* [3] used a diverse range of features i.e. font type, neighbor distance, text location, font typography, and lexical properties to identify components of a scientific publication and later extract metadata like Authors name, affiliation, email, headings, etc. Boukhers *et al.* [20] proposed an approach in which all text lines are individually classified using a pre-trained random forest model with the probability to be a potential reference line and later uses the format, lexical, semantic and shape features to identify and segment reference strings.

Lafferty *et al.* [66] proposed an advanced approach known as Conditional Random Fields (CRF). CRF is a probabilistic approach for labeling sequence data like reference strings. This labeling includes identifying different parts of a reference string i.e. authors, publication title, year, conference/journal name, etc. Such labeling assists in recognizing a reference string based on its labeled components.

Tkaczyk *et al.* [120] proposed "Content ExtRactor and MINEr (*CERMINE*)" a CRF based system for extracting and mining bibliographic metadata from references in born-digital PDF scientific articles. Freecite [44] computes features from tokenized citation string and then classify that token sequence using trained CRF. Science Parse [112] is a tool based on CRF to identify and extract metadata of references from a document. Matsuoka *et al.* [80] demonstrated the use of lexical features by CRF results to gain an increase in accuracy. Councill *et al.* [33] presented a CRF based package called "*ParsCit*" for reference metadata tagging problem. In which the reference strings were identified from plain text, based on fine-grained heuristics. [33] claims *ParsCit* to be one of the best known and widely used open-source system based on Heuristics and CRF for reference detection, string parsing, and metadata tagging. Tkaczyk *et al.* [119] also proposed a reference metadata recommender system which provided 10 most popular open-source citation parser tools in one system. Selected tools were a mixture of simple heuristics based and machine learning-based solutions.

Nowadays, artificial neural networks are the most popular choice as a solution to most scientific problems. Similarly, some literature also explored the potential of neural networks for the task of bibliographic reference detection and parsing.

Zou *et al.* [133] proposed a two steps approach to locate and parse bibliographic references in HTML medical articles. In the first step individual references are located using machine learning approaches whereas in the second step by employing CRFs, metadata is extracted from each reference.

Contrary to the traditional approaches for reference tagging, Parsad *et al.* [99] proposed a bibliographic reference string parser named "*Neural-ParsCit*" based on deep neural networks. The authors tried to capture long-range dependencies in reference strings using Long

Table 2.1: Overall distribution of *BibX*[14] dataset

	Train Set	Validation Set	Test Set
No. of Images	287	25	143
No. of References	5741	478	2547
Single Column	270	24	136
Double Column	17	1	7

Short Term Memory (LSTM) [56] based architecture. Lopez *et al.* [76] proposed a tool named "*Grobid*" based a tool based on conditional random fields for detection and extraction of publication headers, bibliographic references and their respective metadata. The *Grobid* model was trained on multi-domain, manually annotated data containing 6835 instances. Recently, Grennan *et al.* [49] performed experiments to train a CRF-based solution [76] on actual citation parsing data annotated by humans and synthetic data and suggested that the model trained on synthetic dataset performed very similar to the model trained on original data.

Text-based approaches are not directly applicable to document images. To identify references from scanned documents, the text must be extracted from a given document by performing Optical Character Recognition (OCR) and then applying the selected approach to the extracted text. The disadvantage of this approach is the potential introduction of OCR error which will eventually contribute to detection error thus making the task unnecessarily complicated.

2.1.2 *Layout-based Approaches*

The literature discussed so far relies only on textual features to identify references. Text-based approaches do not take advantage of layout features thus abandoning an important aspect. There are very few approaches that explored the potential of exploiting layout information for detecting bibliographic references.

Bhardwaj *et al.* [15] used layout information to detect references from a scanned document. For that purpose, Fully Convolutional Neural Network (FCN) [75] was used to segment the references and later post-processed to identify individual references. To our best knowledge, it is currently the state-of-the-art for the image-based reference detection task. The authors also released a small dataset [14] for image-based reference detection. In this chapter, this dataset will be referred to as *BibX* dataset. Lauscher *et al.* [68] used this layout based reference detection in their system [72] to build an open database of citations for libraries indexing use case. Recently, Rizvi *et al.* [102] gauged the performance of four state-of-the-art object detection

Table 2.2: Overall distribution of *BibLy* dataset

	Train Set	Validation Set	Test Set
No. of Images	1513	132	756
No. of References	24606	2013	12244
Single Column	1411	124	705
Double Column	92	7	46
Triple Column	10	1	5

models using layout information to detect bibliographic references in a scientific publication. [103]

2.2 DATASETS

2.2.1 *BibX dataset*

This section provides insights into the dataset used for training and baseline performance comparison of *DeepBIBX* [15] and our proposed approach *DeepBiRD* [103] for the task of layout-based reference detection. To the best of the authors' knowledge, it is the only image-based dataset that contains annotations of references. *BibX* dataset consists of 455 document images from several social sciences books and journals, containing 429 and 25 document image samples from single and double column layouts respectively. The dataset is divided into train, validation, and test set with 287, 25, and 143 samples respectively. Distribution details of *BibX* dataset are mentioned in Table 2.1. Furthermore, considering the limited size of the dataset, we propose a new dataset called *BibLy* dataset. Details of this new dataset are discussed in the following section.

2.2.2 *BibLy dataset*

In this work, we released a dataset named *BibLy* [38] for image-based reference detection. This dataset has been curated from the reference section of various Journals, Monographs, Articles, and Books from the social sciences domain. The resolution of images varies from 1500 to 4500 for the larger side of the image. Image quality is maintained on at least 300 dpi. All images were manually annotated where a box was drawn around every single reference.

There are 2,401 scanned document images in *BibLy* dataset containing 38,863 references in total. Document scans were initially divided into three groups based on the number of columns i.e. single, double, and triple columns. Table 2.2 shows the distribution of samples in layout groups. These groups were further distributed into train, val-

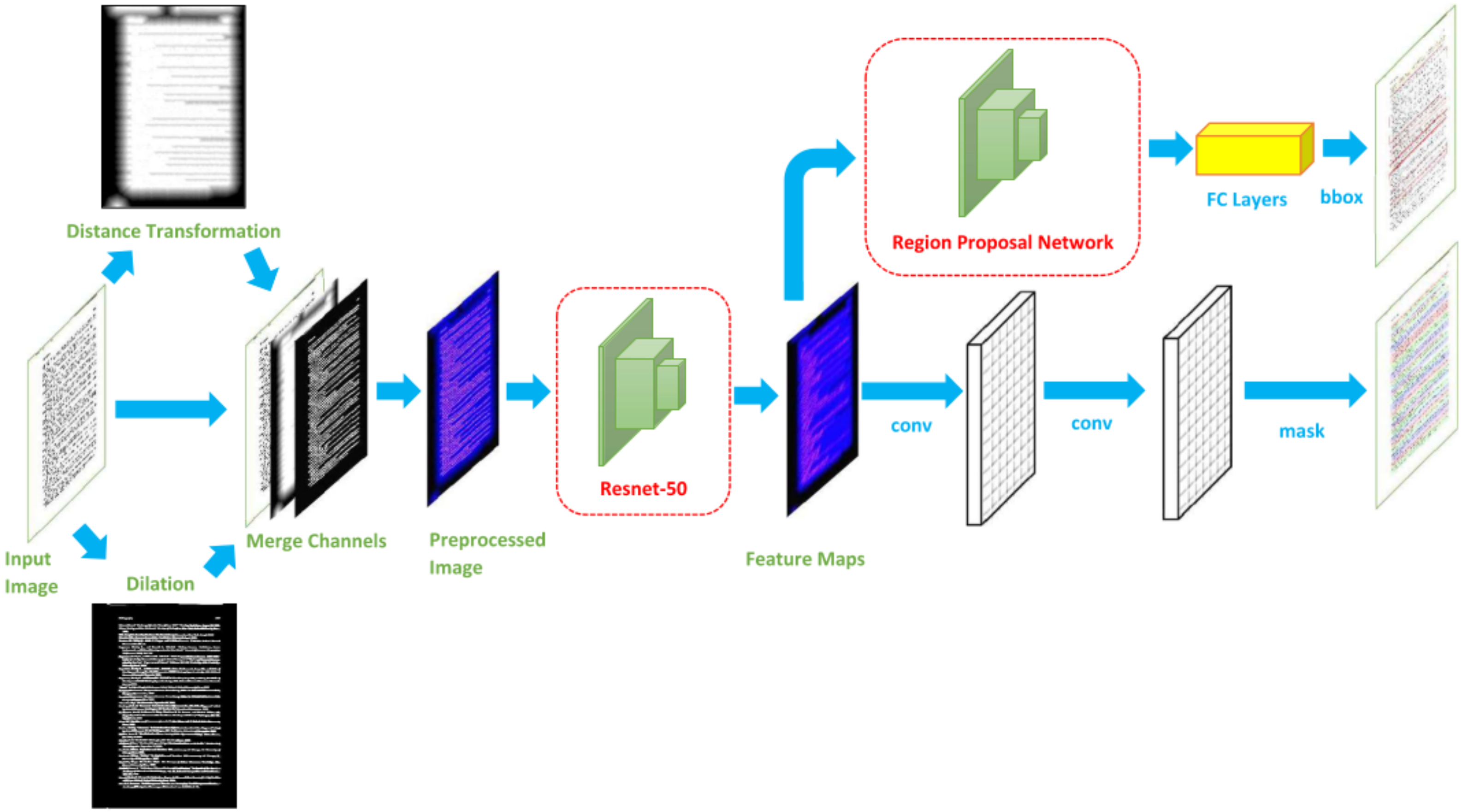
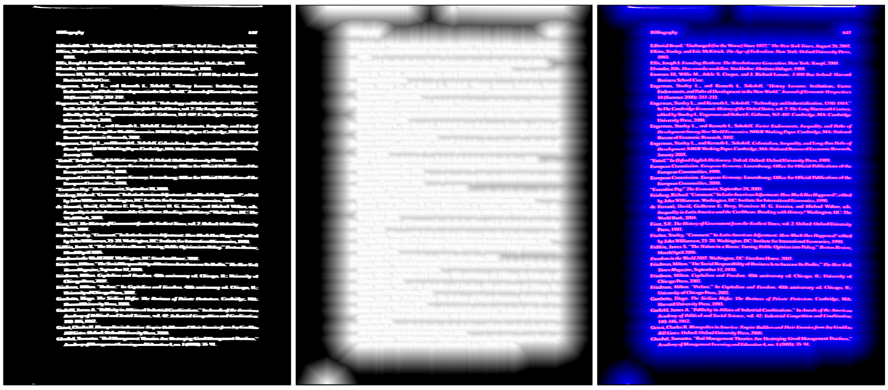


Figure 2.3: Overview of proposed *DeepBiRD* [103] pipeline for layout-based reference detection



(a) Dilated (b) Distance Transform (c) Hybrid Representation

Figure 2.4: Visual examples of a document in different pre-processing stages

idation, and test set with balanced representation from each group. Fig 2.2 shows sample scans with different layouts. The distribution of train test and validation set along with their respective number of references are shown in Table 2.2. Dataset is shared on the following link: `https://madata.bib.uni-mannheim.de/283/`.

2.3 DEEPBIRD: PROPOSED APPROACH

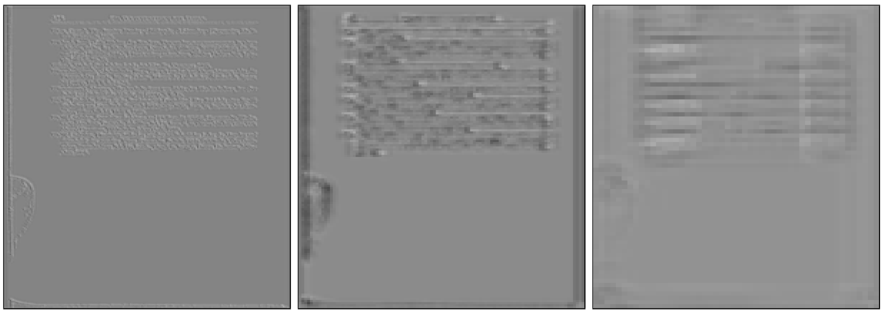
In our proposed approach, we exploit layout information to detect references from a given document image. Firstly, we pre-process the input document image and incorporate more layout information to facilitate bibliographic reference detection. Later, references are detected from each pre-processed image. Fig 2.3 depicts the complete pipeline of our proposed system. Details of our pipeline are discussed as follows.

2.3.1 *Pre-processing*

The first stage in our pipeline is pre-processing followed by reference detection. In order to highlight layout features, we obtained a hybrid representation, which highlights the important content and helps the automatic representation learning approach to extract discriminative features. This hybrid representation is achieved by applying different transformations to the input image. The pre-processing stage involves a series of steps, which are elaborated as follows:

2.3.1.1 *Distance Transform*

Distance transform provides the distance between each pixel and the nearest input foreground pixel. This way, we can highlight the separation between words, lines, and characters which later proves to help identify and separate individual references.



(a) Upscaled conv2 (b) Upscaled conv3 (c) Upscaled conv4

Figure 2.5: Feature maps from different stages in the architecture backbone

To apply distance transform, we first read the image as a grayscale image followed by the inversion of the image therefore switching bright pixels with dark pixels and vice versa. Then we binarize the input image using OTSU thresholding followed by inversion of all pixel values. Distance transform is then applied to the resultant binarized inverted image. We used several distance types in different experiments and selected Euclidean distance with a 3×3 mask as the most suitable distance measure. An example of Euclidean distance transformation is shown in Fig 2.4b.

2.3.1.2 *Dilation*

We performed dilation on the input image to highlight text regions along with their surroundings to facilitate the neural network to identify lines and their respective scope more precisely. To perform dilation, we firstly binarized the input image using OTSU thresholding followed by the inversion. Then we perform dilation using a kernel of 1×5, this horizontal kernel merges the nearby characters in the proximity of the same line. The motivation of using a kernel of 1×5 is to preserve line separation while merging the words in the same line, therefore highlighting a line. A sample image is shown in Fig 2.4a.

2.3.1.3 *Hybrid Representation*

It is the final stage of the pre-processing phase, in which we merge the representations obtained from dilation and distance transform with the input image. For that purpose, we place distance transform image, binarized image, and dilated image in channels one, two, and three of an image respectively. The resultant image retains information of the original image along with additional highlighted text lines and proximity information encoded into one image. In the hybrid representation, the blue color represents the proximity of the text and the separation between words and lines. While red color represents the

Table 2.3: Architecture details of the employed feature backbone ResNet-50 [53]

Layer	Output	Structure
conv1	112×112	7×7, 64, stride2
conv2_x	56×56	3×3, maxpool, stride2 $\begin{bmatrix} 1 \times 1 \times 64 \\ 3 \times 3 \times 64 \\ 1 \times 1 \times 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 1 \times 1 \times 128 \\ 3 \times 3 \times 128 \\ 1 \times 1 \times 512 \end{bmatrix} \times 4$
conv4_x	14×14	$\begin{bmatrix} 1 \times 1 \times 256 \\ 3 \times 3 \times 256 \\ 1 \times 1 \times 1024 \end{bmatrix} \times 6$
conv5_x	7×7	$\begin{bmatrix} 1 \times 1 \times 512 \\ 3 \times 3 \times 512 \\ 1 \times 1 \times 2048 \end{bmatrix} \times 3$
	1×1	averagepool, fc, softmax

color of the line. This image is later used to identify bibliographic references from a given document image. An example of the final image is shown in Fig 2.4c.

2.3.2 *Reference Detection Model*

This section provides insights into the architecture design of the *DeepBiRD* [103].

2.3.2.1 *Architecture*

For the reference detection task, due to the proximity of references a network was needed which can also separate references from each other in addition to detecting those references. For that purpose, we employed a deep neural network-based architecture known as Mask R-CNN [52]. It is one of the most popular networks for object detection and instance segmentation. The task of reference detection using layout features is a challenging task, as the references are a few pixels apart. So the detection task requires high precision to the level of each pixel. In contrast to the Faster-RCNN[101], the Mask R-CNN [52] is equipped with the ROIAlign which is a Non-quantized operation and therefore preserves the data. This resulted in more accurate detections to the pixel level. It served as one of the main reason to employ Mask R-CNN[52] for the task at hand.

Table 2.4: Detection results from all variations of *DeepBiRD* [103] and *DeepBIBX* [15] on both datasets

	Model	Trained on	Tested on	mAP[0.5:0.95]	AP50	AP75	AR
Experiment 1	DeepBiRD	BibX	BibX	76.52	97.59	88.52	80.40
	DeepBIBX [15]	BibX	BibX	32.51	54.22	36.24	23.28
Experiment 2	DeepBiRD	BibX	BibLy	64.53	89.40	75.20	70.50
	DeepBIBX [15]	BibX	BibLy	29.03	52.56	30.27	21.44
Experiment 3	DeepBiRD	BibX	BibLy	64.53	89.40	75.20	70.50
	DeepBiRD	BibX + BibLy	BibLy	83.40	98.56	95.39	86.60

In our experiments we used standard ResNet-50 [53] backbone for feature extraction. Table 2.3 shows the details of ResNet-50 [53] . Following the original implementation in [46], we followed the original parameters of [52] to train the network with batch normalization enabled. We used the pre-trained model ResNet-50 [53] to initialize the network. Then it was fine-tuned on the train set of *BibX* dataset with 287 images containing 5741 references, using transfer learning. For fine-tuning, we froze the first two blocks conv1 and conv2_x while leaving all the remaining blocks trainable. Figure 2.5 shows samples of feature maps from different feature extraction layers in the architecture.

2.3.2.2 *Parameters*

The network was trained for 50 epochs with a base learning rate of 0.001. The learning rate was decreased in steps by a factor of 0.0001 at 12, 25, and 37 epochs respectively. In all experiments, the number of images per batch was set to 1.

2.3.2.3 *Inference*

By performing inference on the input image we get coordinates of the detected reference's box along with a confidence score. The confidence score ranges between 0 and 1, where 0 being lowest and 1 being highest. It represents the extent to which the network is sure about that specific detection. Each detection in the results represents a reference. Once detection results are ready, OCR[1] is performed on each detected reference, thus extracting all references from an input image.

2.4 EXPERIMENTS & RESULTS

To evaluate our system, we performed various experiments using *DeepBIBX* [15] model and multiple settings of *DeepBiRD* [103] on two publicly available datasets. These datasets include our own *BibLy* [38]

[1] https://github.com/tesseract-ocr/tesseract

Table 2.5: Results from ablation study of *DeepBiRD* [103] on both datasets

Dataset	Pre-processing Type	mAP[0.5:0.95]	AP50	AP75	AR
BibX	Dilation + Distance Transform	76.52	97.59	88.52	80.40
	Dilation	75.91	96.90	88.29	79.80
	No Pre-processing	75.28	96.85	87.32	79.40
BibLy	Dilation + Distance Transform	83.40	98.56	95.39	86.60
	Dilation	83.24	98.54	95.35	86.50
	No Pre-processing	83.16	98.50	95.30	86.30

dataset and the one proposed by *DeepBIBX* [15], here referred to as *BibX* dataset [14]. Due to the limited number of samples in *BibX*, the authors augmented the whole dataset followed by resizing every image in train, validation and test set. In this section, we will elaborate the results of the experiments performed for evaluation.

2.4.1 *Evaluation*

In this section, we will discuss the evaluation results of all the experiments performed to compare *DeepBiRD* [103] model with *DeepBIBX* in different settings.

2.4.1.1 *Experiment 1: Baseline comparison of our approach with DeepBIBX [15]*

The purpose of this experiment was to validate the effectiveness of our approach on *BibX* dataset and compare its performance with *DeepBIBX* [15]. In this experiment, we trained *DeepBiRD* [103] on *BibX* dataset with aforementioned parameters. We also trained a Fully Convolutional Network (FCN) [75] on non-augmented *BibX* dataset with exactly same settings as mentioned in the *DeepBIBX* original paper [15]. The difference being that in our experiments we used non-augmented dataset to enable results to be directly comparable with our approach. Additionally, we resized the training, validation or test set images and blurred its lines as mentioned in [15]. However, It is worth mentioning that using non-augmented dataset, will result in different evaluation results from *DeepBIBX* [15]. Once the training finished, both models were evaluated on non-augmented test set of *BibX* dataset. By doing so it enabled us to directly compare the performance of our approach with *DeepBIBX* [15] approach on *BibX* dataset.

Each detection was validated using its Intersection over Union (IoU) with ground truth annotations. Both models were evaluated on different IoU thresholds ranging from 0.50 to 0.95 which is a standard for an object detection problem. A detection is considered as correct detection if the IoU of a given bounding box is greater than the IoU threshold. Table 2.4 shows comparison of *DeepBiRD* [103] results with

DeepBIBX [15] for experiment 1. The results show that *DeepBiRD* [103] was able to achieve an average precision and average recall of 76.52% and 80.40% respectively. On the other hand *DeepBIBX* was only able to achieve an average precision and average recall of 32.51% and 23.28% respectively. Even at the lowest IoU threshold of 0.50, *DeepBiRD* [103] was able to perform significantly better even more than a factor of 2.

The reason behind the strong performance of *DeepBiRD* [103] is that it is based on Mask R-CNN [52] which performs semantic segmentation on shortlisted ROIs on the other hand FCN performs semantic segmentation on complete image.

2.4.1.2 *Experiment 2: Robustness*

The purpose of this experiment was to validate the extent of robustness for both *DeepBiRD* [103] and *DeepBIBX* [15]. To do so, we evaluated both systems on more unseen data i.e. test set from another dataset. The *DeepBiRD* [103] and *DeepBIBX* models trained in the Experiment 1 were reused in this experiment. Both models were trained on Non-Augmented *BibX* dataset and evaluated on test set of *BibLy* dataset. The results from this experiment show the extent of effectiveness of *DeepBIBX* [15] & *DeepBiRD* [103] on unseen data. Both models were evaluated on a range of IoU thresholds ranging from 0.50 to 0.95. Table 2.4 shows the evaluation results of *DeepBIBX* [15] model on *BibLy* dataset. The results show that the performance of both models slightly decreased as expected when they are applied to unseen data. *DeepBiRD* [103] was able to achieve an average precision and average recall of 64.53% and 70.50% respectively. On the other, *DeepBIBX* was able to achieve an average precision and average recall of 29.03% and 21.44% respectively. Therefore, outperforming *DeepBIBX* by a significant margin similar to experiment 1 results.

2.4.1.3 *Experiment 3: Generalization*

The purpose of this experiment was to verify *DeepBiRD* [103] for generalization by employing transfer learning to adapt the network to the *BibLy* dataset. In this experiment, the pre-trained *DeepBiRD* [103] model on *BibX* dataset was used as a baseline and was then fine-tuned on the train set of *BibLy* dataset to learn more reference examples. Once the training was finished, the final model was evaluated against the baseline model on the test set of *BibLy* dataset.

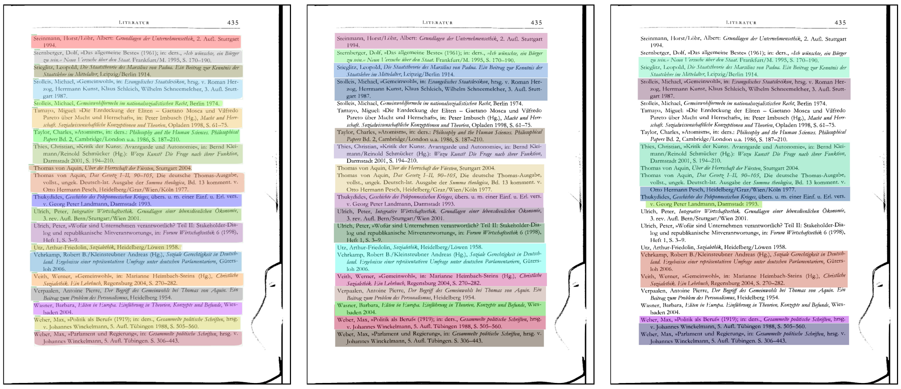
The results of this experiment are shown in Table 2.4. The fine-tuned model was able to achieve an average precision and average recall of 83.40% and 86.60% respectively, which is 18.87% and 16.1% better than the results before fine-tuning of the model. However, precision at IoU of 0.50 and 0.75 increased by 9.16% and 20.19% respectively. This indicates that after fine-tuning, the model was significantly improved and was able to detect bibliographic references with a higher overlap. From these results, we can infer that *DeepBiRD* [103] can be generalized as it can adapt very well to new data.

2.4.2 *Ablation Study for Input Representation*

This section discusses the results of the ablation study to show the impact of hybrid representation used in *DeepBiRD* [103]. The purpose of this analysis was to determine the effectiveness of individual components in the pre-processing phase. For this analysis, we designed several experiments with different pre-processing configurations. In the first experiment, we employed the aforementioned pre-processing steps i.e. distance transform, dilation, and merging them with the original input image. In the second experiment, we employed dilation as a sole step in the pre-processing phase. Lastly in the third experiment, we excluded the pre-processing phase and used the original input image without any pre-processing. These experiments will highlight the contribution of individual components in the pre-processing phase towards the final output.

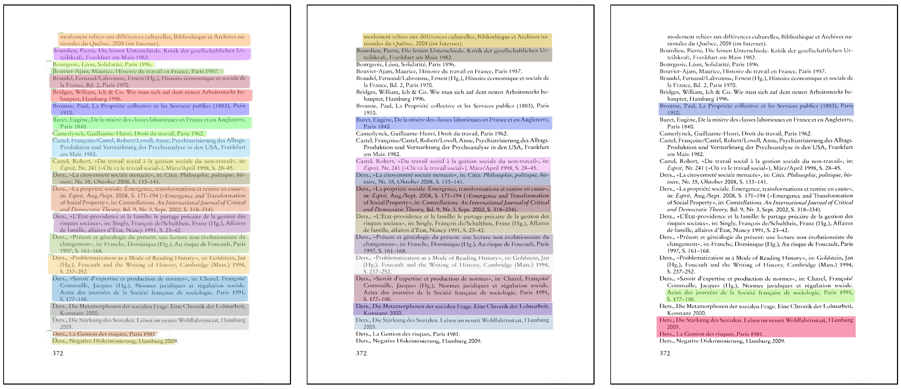
We used *BibX* dataset to perform this ablation study. Table 2.5 shows the results of different representation types employed in various experiments. The evaluation results show that the experiment which employed both dilation and distance transform along with merging channels sets the baseline average precision and an average recall of 76.52% and 80.40% respectively. In the second experiment, pre-processing consisted of dilation of the input image. This resulted in a decrease of 0.6% in both average precision and average recall, therefore suggesting that providing distance transform aided the proposed system to detect bibliographic references from a given document image. In the third experiment, the pre-processing phase was removed altogether and the original input image was fed to the system with no pre-processing. This resulted in a further decrease in average precision and average recall by 0.63% and 0.40% respectively. Therefore suggesting that dilation also contributed towards improving system performance.

To verify the trend in results, we performed the same ablation study on a second dataset *BibLy* dataset. Table 2.5 shows the results of this analysis. In the first experiment, with dilation and distance transform as a part of pre-processing, it sets the baseline evaluation average precision and an average recall of 83.40% and 86.60% respectively. The second experiment with pre-processing involving dilation of the input image decreased the average precision and average recall by 0.16% and 0.10% respectively. Whereas in the third experiment with no pre-processing, the average precision and average recall were further decreased by 0.08% and 0.20% respectively. These trends in re-



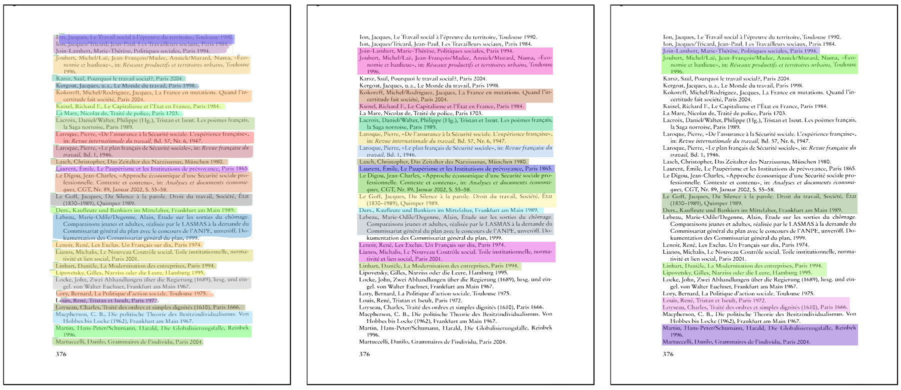
(a) DeepBiRD (b) DeepBIBX[15] (c) ParsCit[33]

Figure 2.6: Best case output sample in comparison with state-of-the-art approaches



(a) DeepBiRD (b) DeepBIBX[15] (c) ParsCit[33]

Figure 2.7: Average case output sample in comparison with state-of-the-art approaches



(a) DeepBiRD (b) DeepBIBX[15] (c) ParsCit[33]

Figure 2.8: Worst case output sample in comparison with state-of-the-art approaches

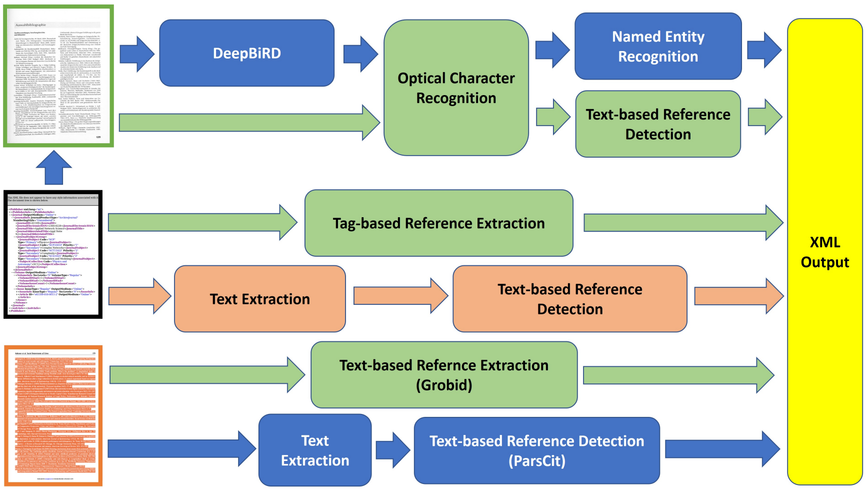

DeepBiRD
Optical Character Recognition
Named Entity Recognition
Text-based Reference Detection
Tag-based Reference Extraction
Text Extraction
Text-based Reference Detection
Text-based Refernce Extraction (Grobid)
Text Extraction
Text-based Reference Detection (ParsCit)
XML Output

Figure 2.9: Overview of the BRExSys

sults proved to be consistent that both dilation and distance transform play their part in further improving the performance of the system.

Fig 2.6, 2.7 & 2.8 show visual examples of best, average and worst results from our system. Results from *DeepBIBX* [15] and text-based model *ParsCit* for each example are also shown for comparison. All these results demonstrate the dominance of *DeepBiRD* [103] over all other text-based or layout-based approaches. Trained model of the above mentioned experiments is available at the URL[2].

2.5 BREXSYS: A BIBLIOGRAPHIC REFERENCE EXTRACTION SYSTEM

This section discusses the details of our proposed framework. Our framework unifies all state-of-the-art bibliographic reference detection methods in one place to detect and extract references from scanned, markup, and textual documents. To take advantage of multiple models to the full extent, we provide various possibilities to use these models individually or in a fusion. The overview of our complete system is shown in Fig 2.9 and details of the proposed system are discussed as follows:

[2] https://github.com/rtahseen/DeepBiRD

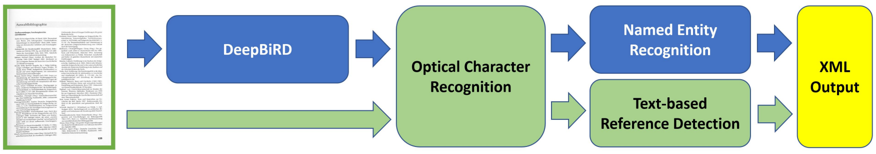


Figure 2.10: Pipeline for Scanned Documents

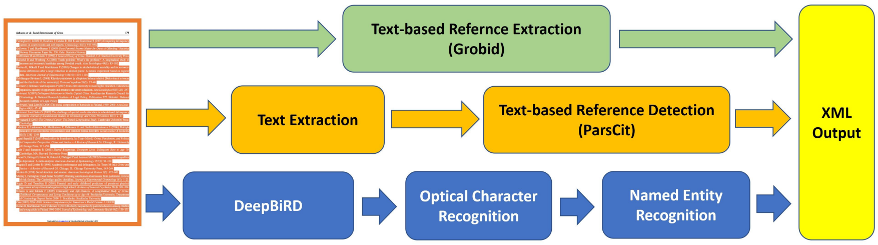

Text-based Refernce Extraction (Grobid)
Text Extraction
Text-based Reference Detection (ParsCit)
DeepBiRD
Optical Character Recognition
Named Entity Recognition
XML Output

Figure 2.11: Pipeline for Textual Documents

2.5.1 *Reference Extraction from Scanned Documents*

In this section, we will discuss pipelines specific for bibliographic reference extraction from scanned documents. The overview of these pipelines is shown in Fig 2.10. For scanned documents, we provide two pipelines i.e. Layout-based pipeline and Text-based pipeline. The layout-based pipeline is represented in blue color while the text-based pipeline is represented in green color. A scanned document can also be processed through both pipelines simultaneously and for such cases, results from both pipelines are included in the final output XML.

2.5.1.1 *Layout-based pipeline*

In a layout-based pipeline, we employed *DeepBiRD* [103], a state-of-the-art layout-driven reference extraction model. Provided a scanned document *DeepBiRD* [103] performs bibliographic reference detection on the individual document image. Lastly, we employed *ParsCit* [33] to carry out Named Entity Recognition (NER) on each detected reference to identify reference string metadata like author names, publication title, publication year, etc. All the results are eventually returned in the form of a predefined standard XML file format.

2.5.1.2 *Text-based pipeline*

In text-based pipeline, we extract all the text from given scanned document and use it for text-based reference extraction. For this purpose we employed *ParsCit* [33], a state-of-the-art text-based reference extraction model. Additionally, *ParsCit* [33] extracts reference string metadata by performing NER on extracted bibliographic reference strings.

2.5.2 *Reference Extraction from Textual Documents*

This section discusses reference extraction pipelines from text documents like born-digital PDFs and plain text files. The pipeline overview for bibliographic reference extraction from textual documents is shown

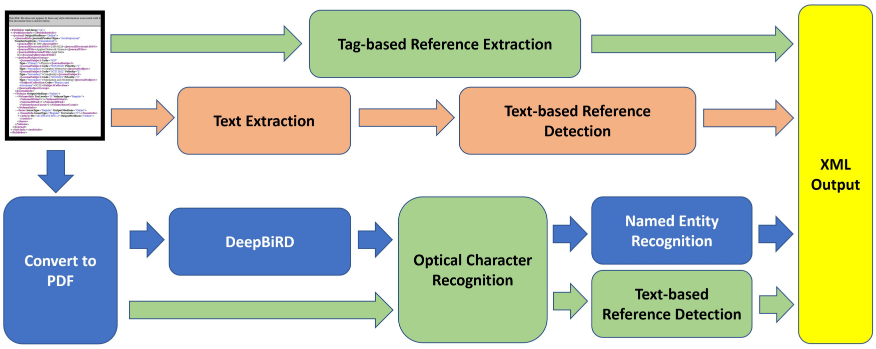

Tag-based Reference Extraction
Text Extraction
Text-based Reference Detection
XML Output
Convert to PDF
DeepBiRD
Optical Character Recognition
Named Entity Recognition
Text-based Reference Detection

Figure 2.12: Pipeline for Markup Documents

in Fig 2.11. We provide three pipelines for extracting references and their respective metadata from a given textual document. Each of these pipelines is discussed as follows:

2.5.2.1 *Text-based pipeline (Grobid)*

In this pipeline, we employed *Grobid* [76]. It takes born-digital PDF as an input and extracts bibliographic references along with their metadata from a given PDF document. Extracted data is returned in the form of predefined XML. *Grobid* does not depend upon other tools for text extraction from a given PDF document, therefore avoiding the introduction of a potential text extraction error.

2.5.2.2 *Text-based pipeline (ParsCit)*

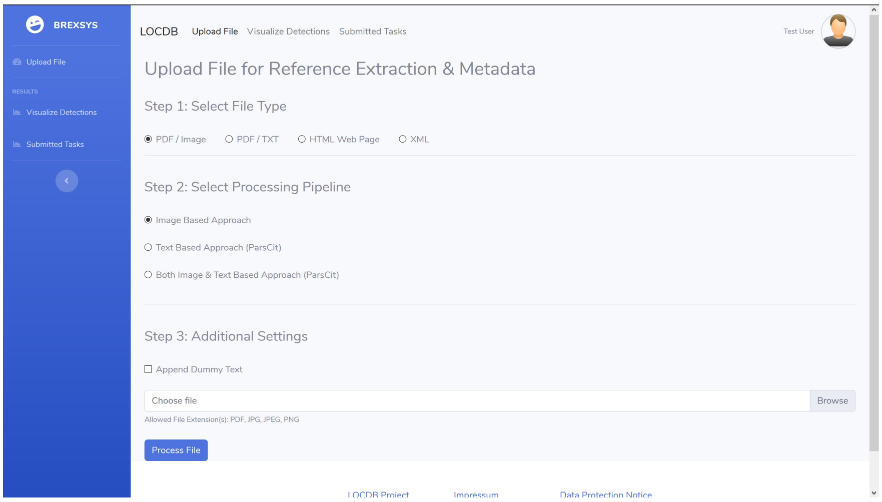
In this pipeline, we firstly extract text from a given textual document and is further processed for bibliographic reference and metadata extraction. For this workflow we employed *ParsCit* [33]. It takes raw text as an input and extracts references along with its metadata from the given text.

2.5.2.3 *Layout-based pipeline*

Tertiary workflow serves as an alternate solution suggesting that a born-digital PDF can also be processed as a scanned document using layout-driven reference extraction. In this workflow, we employed a state-of-the-art layout based reference detection and extraction approach called *DeepBiRD* [103]. Details of this pipeline are already discussed in the section discussing reference extraction from Scanned documents.

2.5.3 *Reference extraction from markup documents*

In this section, we will discuss bibliographic reference extraction from markup documents like HTML and XML. Markup documents usu-



Upload File for Reference Extraction & Metadata

Step 1: Select File Type

● PDF / Image ○ PDF / TXT ○ HTML Web Page ○ XML

Step 2: Select Processing Pipeline

● Image Based Approach

○ Text Based Approach (ParsCit)

○ Both Image & Text Based Approach (ParsCit)

Step 3: Additional Settings

□ Append Dummy Text

Choose file

Allowed File Extension(s): PDF, JPG, JPEG, PNG

Figure 2.13: Input interface

ally consist of a peculiar hierarchy. Depending on a known topology of a given markup document we provide multiple workflows for extracting bibliographic references and their metadata from markup documents. The overview of the pipelines is shown in Fig 2.12, where each pipeline handling a specific case is represented in a different color.

2.5.3.1 *Direct mapping pipeline*

The direct mapping pipeline deals with the case when we are fully aware of tags hierarchy in the markup document i.e. XML or HTML document from Zotero [132]. In such cases we perform tag-based reference extraction by targeting relevant tags like author name, title, publisher, etc., thus extracting all references from markup documents along with their metadata.

2.5.3.2 *Text-based pipeline*

Text-based pipeline deals with the case where we have partial knowledge about the tags hierarchy of a given markup document i.e. XML or HTML document generated from older versions of Zotero [132]. In such cases, we first extract all the text from the markup document using all known tags and then perform text-based reference extraction on extracted text. We employ *ParsCit* [33] for extracting references and their respective metadata from the extracted text.

2.5.3.3 *Layout-based pipeline*

Layout-based pipeline deals with the case when a given markup document is in HTML format and has an unknown tags hierarchy. In this case, we will convert the HTML document into a PDF document and

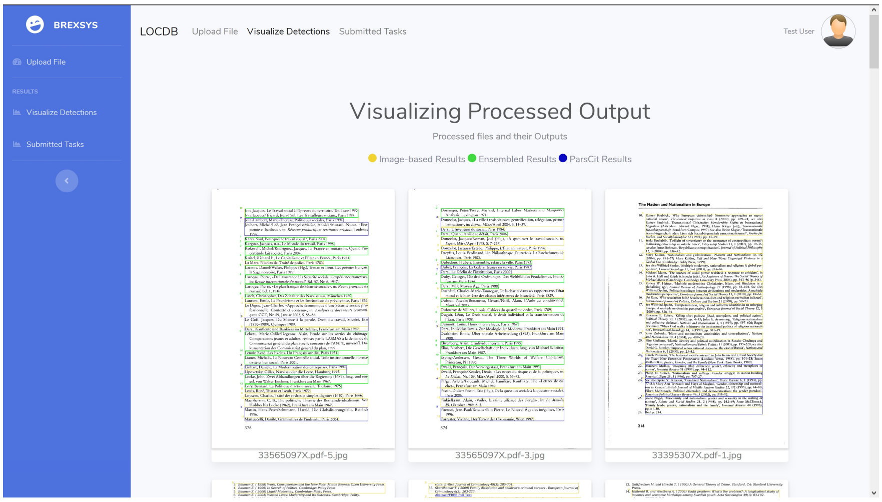


Figure 2.14: Output visualizing interface

process it as a scanned PDF where it is simultaneously processed using text and layout-based bibliographic reference extraction pipelines.

2.5.4 *Interfaces and Output*

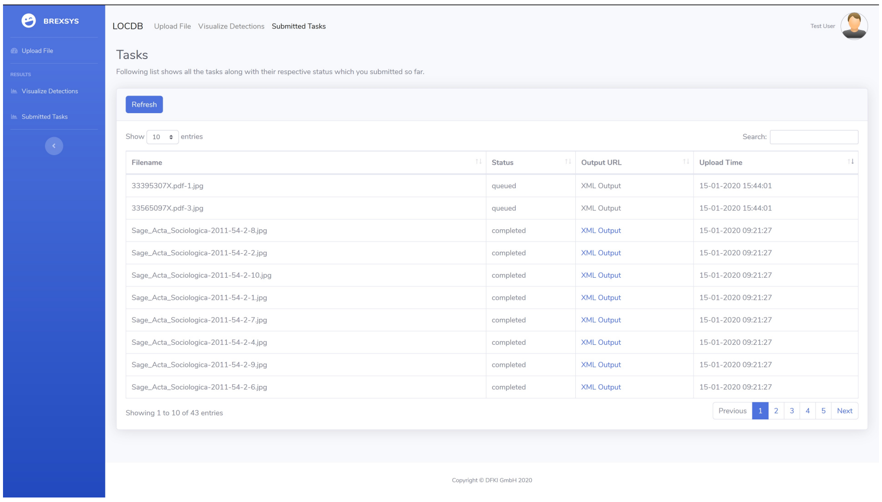
In this section, we will discuss different interfaces and outputs sample our of the proposed system. Our system provides a web-based friendly interface where one interface for uploading and configuring files for processing while the other interface is responsible for displaying the results from layout-based detection. Additionally, an interface also lists all submitted processing tasks along with their output.

2.5.4.1 *Input interface*

The input interface of our system is shown in Fig 2.13. User can upload any file type with extension PDF, JPG, PNG, TIF, TXT, HTML and XML. In the first step, the user selects the desired file type for processing. Once the file type is selected, all available relevant pipelines are revealed. After selecting the desired processing pipeline, the user is asked to upload the desired file. Additionally, users can check an additional option on whether or not to add dummy text before the extracted text. For this purpose "**Append Dummy Text**" flag must be enabled. During the evaluation of *ParsCit*[33] we found out that appending dummy text to the start of the references text yields better results. Once all settings are done the user can trigger the processing phase by pressing **Process File** button.

2.5.4.2 *Output visualizing interface*

Output visualizing interface is another important interface of our system where we can visualize the results of all documents processed as



Tasks

Following list shows all the tasks along with their respective status which you submitted so far.

Show 10 entries

Search:

Filename	Status	Output URL	Upload Time
33395307X.pdf-1.jpg	queued	XML Output	15-01-2020 15:44:01
33565097X.pdf-3.jpg	queued	XML Output	15-01-2020 15:44:01
Sage_Acta_Sociologica-2011-54-2-8.jpg	completed	XML Output	15-01-2020 09:21:27
Sage_Acta_Sociologica-2011-54-2-2.jpg	completed	XML Output	15-01-2020 09:21:27
Sage_Acta_Sociologica-2011-54-2-10.jpg	completed	XML Output	15-01-2020 09:21:27
Sage_Acta_Sociologica-2011-54-2-1.jpg	completed	XML Output	15-01-2020 09:21:27
Sage_Acta_Sociologica-2011-54-2-7.jpg	completed	XML Output	15-01-2020 09:21:27
Sage_Acta_Sociologica-2011-54-2-4.jpg	completed	XML Output	15-01-2020 09:21:27
Sage_Acta_Sociologica-2011-54-2-9.jpg	completed	XML Output	15-01-2020 09:21:27
Sage_Acta_Sociologica-2011-54-2-6.jpg	completed	XML Output	15-01-2020 09:21:27

Showing 1 to 10 of 43 entries

Copyright © DFKI GmbH 2020

Figure 2.15: Tasks status interface

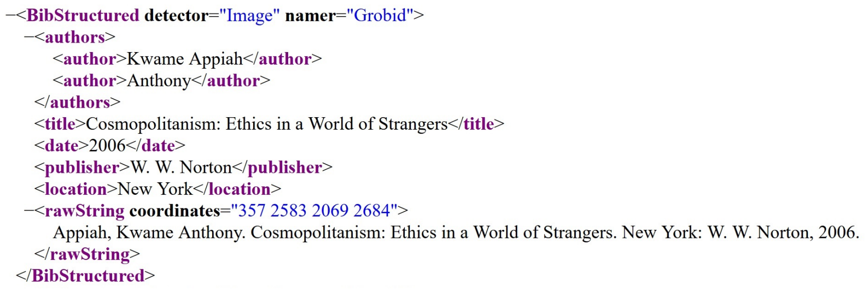
scanned documents. Fig 2.14 shows the output visualizing interface of our system containing all the images/scanned PDF documents already process. The detected references from both layout and text-based models are represented in different colors. Results from the layout-based model and text-based model are represented in yellow and blue respectively. While the boxes in green represent the references detected by both layout and text-based models.

2.5.4.3 *Tasks status interface*

This interface provides a list of all submitted processing tasks along with their history. Additionally, it also shows their current status whether a task is currently in the queue or is already processed. Fig 2.15 shows the screenshot of the interface. Link to the XML output of each processed file is also available in front of each filename. It is to be noted that user access is protected with logins and sessions, therefore a user will only be able to view their processing tasks.

2.5.4.4 *XML Output*

Our system combines the output from all pipelines for the respective file type and returns it in a single XML file where results from each model are differentiated using two custom XML attributes. Fig 2.16 and 2.17 show output samples from layout and text-based model respectively. The custom attributes are added to identify the source of the output. First attribute is *detector* which refers to the approach used to detected references i.e. image-based or *ParsCit*. The second attribute is *namer* which refers to the approach used to extract reference metadata i.e. author names, title, publisher, etc from raw reference string. The possible values for *namer* are either *ParsCit* or *Grobid*.



```
–<BibStructured detector="Image" namer="Grobid">
  –<authors>
      <author>Kwame Appiah</author>
      <author>Anthony</author>
    </authors>
    <title>Cosmopolitanism: Ethics in a World of Strangers</title>
    <date>2006</date>
    <publisher>W. W. Norton</publisher>
    <location>New York</location>
  –<rawString coordinates="357 2583 2069 2684">
      Appiah, Kwame Anthony. Cosmopolitanism: Ethics in a World of Strangers. New York: W. W. Norton, 2006.
    </rawString>
 </BibStructured>
```

Figure 2.16: XML output from layout-based model

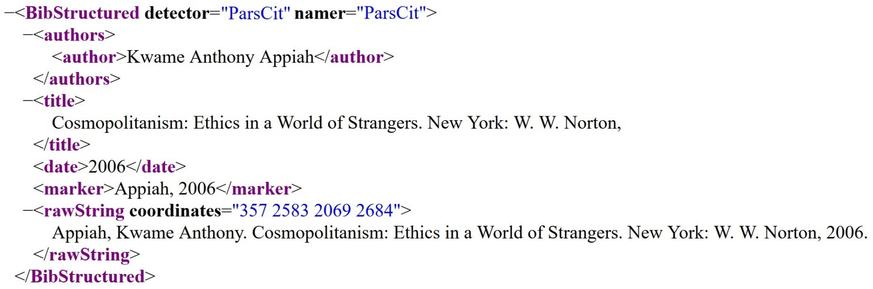
2.5.5 *BRExSys Evaluation*

There are different pipelines in the *BRExSys* framework, where the pipeline with an ensemble of Text and layout-based methods is the largest. There are several phases in the ensemble pipeline. The pre-processing phase takes ≈ 4.35 seconds, followed by reference detection from the image takes ≈ 2.79 seconds which is further followed by the extraction of detections takes ≈ 1.95 seconds. OCR phase takes ≈ 3.63 seconds followed by the most expensive string segmentation phase by image-based approaches which takes ≈ 8.65 seconds. Lastly, compiling both layout and text-based results in an XML file and drawing results on the input image takes ≈ 3.88 and ≈ 1.59 seconds. It is worth mentioning that due to limited resources all these different services were mostly running on a single core which contributed towards using more execution time. *BRExSys* was tested on a system with the following hardware specifications:

- Processor: Intel® Xeon(R) E3-1245 v6 @ 3.70GHz
- Cores: 8
- Graphics: GeForce GTX 1080 Ti
- Memory: 64 GB

To evaluate *BRExSys* on some more challenging cases, we prepared hypothetical examples of different cases using an actual sample from *BibX* dataset. Fig. 2.18 shows the output of all artificially created hypothetical cases after processing them through the ensemble pipeline of layout and text-based models. The output of the original image in Fig. 2.18a serves as the baseline, where all references are perfectly detected. The detections of the ensemble, layout-based, and text-based models are represented in green, yellow, and blue colors respectively.

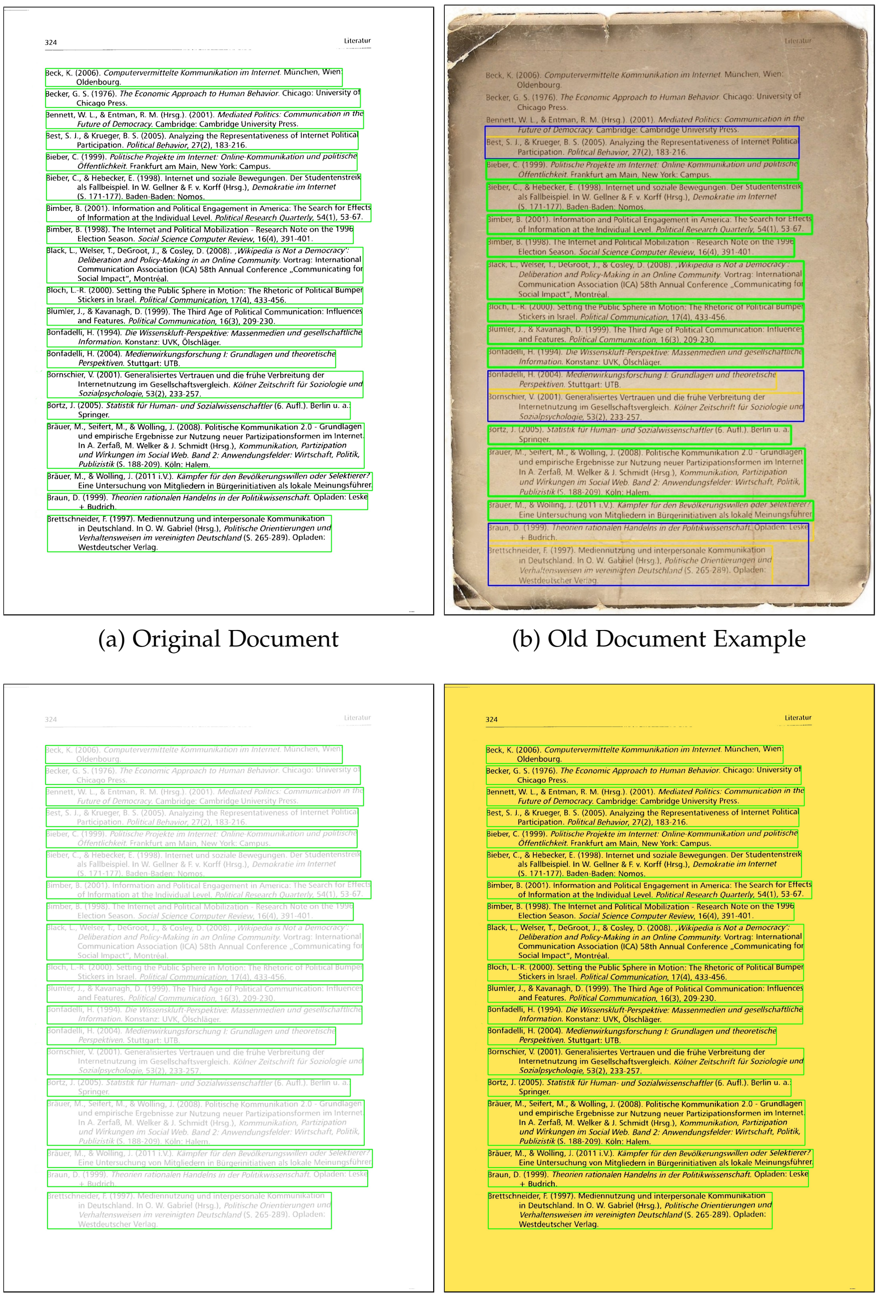
Fig. 2.18b simulates the output of *BRExSys* for an old document. The level of noise in the old document affected the output of the system, where most of the references were successfully detected by the layout-based approach only missing the top three references. Similarly, the text-based model also detected most of the references while



```
−<BibStructured detector="ParsCit" namer="ParsCit">
  −<authors>
      <author>Kwame Anthony Appiah</author>
    </authors>
  −<title>
      Cosmopolitanism: Ethics in a World of Strangers. New York: W. W. Norton,
    </title>
    <date>2006</date>
    <marker>Appiah, 2006</marker>
  −<rawString coordinates="357 2583 2069 2684">
      Appiah, Kwame Anthony. Cosmopolitanism: Ethics in a World of Strangers. New York: W. W. Norton, 2006.
    </rawString>
  </BibStructured>
```

Figure 2.17: XML output from Text-based model

missing the top three references. However, in the case of a text-based model, some of the references are merged and detected as one reference. Fig. 2.18c and Fig. 2.18d simulate the example of dim and tinted document images respectively with different noise levels. However, in both cases *BRExSys* successfully detected all references suggesting that only very high levels of noise may affect the output of the system.



324 Literatur

Beck, K. (2006). *Computervermittelte Kommunikation im Internet.* München, Wien: Oldenbourg.

Becker, G. S. (1976). *The Economic Approach to Human Behavior.* Chicago: University of Chicago Press.

Bennett, W. L., & Entman, R. M. (Hrsg.). (2001). *Mediated Politics: Communication in the Future of Democracy.* Cambridge: Cambridge University Press.

Best, S. J., & Krueger, B. S. (2005). Analyzing the Representativeness of Internet Political Participation. *Political Behavior*, 27(2), 183-216.

Bieber, C. (1999). *Politische Projekte im Internet: Online-Kommunikation und politische Öffentlichkeit.* Frankfurt am Main, New York: Campus.

Bieber, C., & Hebecker, E. (1998). Internet und soziale Bewegungen. Der Studentenstreik als Fallbeispiel. In W. Gellner & F. v. Korff (Hrsg.), *Demokratie im Internet* (S. 171-177). Baden-Baden: Nomos.

Bimber, B. (2001). Information and Political Engagement in America: The Search for Effects of Information at the Individual Level. *Political Research Quarterly*, 54(1), 53-67.

Bimber, B. (1998). The Internet and Political Mobilization - Research Note on the 1996 Election Season. *Social Science Computer Review*, 16(4), 391-401.

Black, L., Welser, T., DeGroot, J., & Cosley, D. (2008). *‚Wikipedia is Not a Democracy': Deliberation and Policy-Making in an Online Community.* Vortrag: International Communication Association (ICA) 58th Annual Conference „Communicating for Social Impact", Montréal.

Bloch, L.-R. (2000). Setting the Public Sphere in Motion: The Rhetoric of Political Bumper Stickers in Israel. *Political Communication*, 17(4), 433-456.

Blumler, J., & Kavanagh, D. (1999). The Third Age of Political Communication: Influences and Features. *Political Communication*, 16(3), 209-230.

Bonfadelli, H. (1994). *Die Wissenskluft-Perspektive: Massenmedien und gesellschaftliche Information.* Konstanz: UVK, Ölschläger.

Bonfadelli, H. (2004). *Medienwirkungsforschung I: Grundlagen und theoretische Perspektiven.* Stuttgart: UTB.

Bornschier, V. (2001). Generalisiertes Vertrauen und die frühe Verbreitung der Internetnutzung im Gesellschaftsvergleich. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 53(2), 233-257.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (6. Aufl.). Berlin u. a.: Springer.

Bräuer, M., Seifert, M., & Wolling, J. (2008). Politische Kommunikation 2.0 - Grundlagen und empirische Ergebnisse zur Nutzung neuer Partizipationsformen im Internet. In A. Zerfaß, M. Welker & J. Schmidt (Hrsg.), *Kommunikation, Partizipation und Wirkungen im Social Web. Band 2: Anwendungsfelder: Wirtschaft, Politik, Publizistik* (S. 188-209). Köln: Halem.

Bräuer, M., & Wolling, J. (2011 i.V.). *Kämpfer für den Bevölkerungswillen oder Selektierer?* Eine Untersuchung von Mitgliedern in Bürgerinitiativen als lokale Meinungsführer.

Braun, D. (1999). *Theorien rationalen Handelns in der Politikwissenschaft.* Opladen: Leske + Budrich.

Brettschneider, F. (1997). Mediennutzung und interpersonale Kommunikation in Deutschland. In O. W. Gabriel (Hrsg.), *Politische Orientierungen und Verhaltensweisen im vereinigten Deutschland* (S. 265-289). Opladen: Westdeutscher Verlag.

(a) Original Document

(b) Old Document Example

(c) Dimmed Document Example

(d) Tinted Document Example

Figure 2.18: Output of BRExSys for artificially created challenging hypothetical cases

3 KEYWORD DETECTION FROM SCIENTIFIC PUBLICATIONS

Keywords are of significant importance as they carry and represent the essence of a text collection. Due to the sheer volume of the available textual data, there has been an increase in demand for reliable keyword detection systems which can automatically, effectively and efficiently detect the best representative words from a given text. Automatic keyword detection is a crucial task for various applications. Some of its renowned applications include information retrieval, text summarization, and topic detection. In a library environment with thousands or millions of literature artifacts, e.g. books, journals or conference proceedings, automatic keyword detection from each scientific artifact [92] can assist in automatic indexing of scientific literature for the purpose of compiling library catalogs.

3.1 LITERATURE REVIEW

In 2014, about 2.5 million scientific articles were published in journals across the world [123]. This increased to more than 3 million articles published in 2018 [61]. It is certainly impractical to manually link huge volumes of scientific publications with appropriate representative keywords. Therefore, a system is imminent which can automatically analyze and index scientific articles. There has been quite a lot of research on the topic of automated keyword detection, however most of the approaches deal with social media like tweets [11, 17, 19, 25, 26, 36, 41, 79, 90, 92, 97, 100].

A popular approach for keyword detection is representing text as an undirected graph $G = (N, E)$, where the nodes N in graph G correspond to the individual terms in the text and the edges E correspond to the relation between these terms. The most popular relation is term co-occurrence, i.e. an edge is added to the graph between nodes n_1 and n_2 if both corresponding terms co-occur within a given sliding window. The recommended window size depends on the selected approach and often lies in the range between 2 and 10 [73, 85, 104]. Duari and Bhatnagar [36] note that the window size w has a strong influence on the properties of the resulting graph. With the increase in w, the density also increases while the average path length between any two nodes decreases.

The assumption behind this sliding window is that the words appearing closer together have some potential relationship [104]. There are several variations of the sliding window, e.g. letting the window slide over individual sentences rather than the entire text and stopping at certain punctuation marks [73]. Duari and Bhatnagar [36] proposed a new concept named Connectivity Aware Graph (CAG): Instead of using a fixed window size, they use a dynamic window size that always spans two consecutive sentences. They argue that consecutive sentences are related to one another. This is closely related to the concept of *pragmatics* i.e. transmission of meaning depending on the context, which is extensively studied in linguistics [27, 62, 84]. In their experiments, they showed that the performance of approaches generally increases when they use CAG instead of graphs built using traditional window sizes.

The first stage comprises of a novel unsupervised keyword detection approach called Collective Connectivity-Aware Node Weight (CoCoNoW). Our proposed approach essentially combines the concepts of Collective Node Weight [17], CAG [36] and Positional Weight [41] to identify, estimate and sort keywords based on their respective weights. We evaluated our approach on three different publicly available datasets containing scientific publications on various topics and with different lengths. The results show that CoCoNoW outperforms other state-of-the-art keyword detection approaches consistently across all three data sets. In the second stage, detected keywords are used in combination with the Computer Science Ontolog (CSO) 3.1[1] [107] to identify topics for individual publications.

The contributions of this chapter are as follows:

- We present a novel graph-based keyword detection approach that identifies representative words from a given text and assigns weights to rank them in the order of relevance.
- We also evaluated our proposed approach on three different publicly available datasets and consistently outperformed all other existing approaches.
- In this work, we also complement our keyword detection system with ontology-based topic modeling to identify topics from a given publication.

3.2 PROPOSED APPROACH

This work proposed a two-staged novel approach in which the first stage deals with automatic keyword detection called CoCoNoW and in the second stage, the detected keywords are consolidated with the Computer Science Ontology [107] to identify topics for a given scientific publication. In CoCoNoW, we present a unique fusion of Collective Node Weight [17], CAG [36] and Positional Weight [41] to identify keywords from a given document in order to cluster publications

[1] https://cso.kmi.open.ac.uk, accessed June-2022

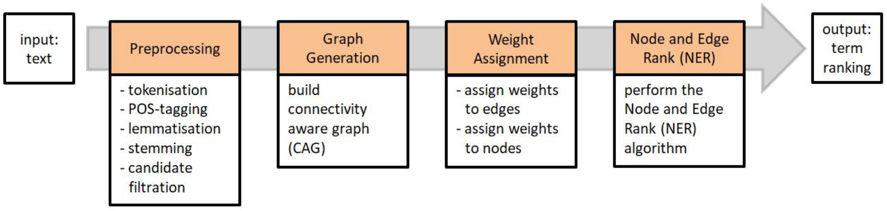


Figure 3.1: An overview of Stage 1 (CoCoNoW) for automatic keyword detection

with common topics together. Fig 3.1 shows the overview of CoCoNoW pipeline. Details of the proposed approach are as follows:

3.2.1 *Stage 1: Automatic keyword detection using CoCoNoW*

3.2.1.1 *Preprocessing*

CoCoNoW uses the standard preprocessing steps like tokenization, part of speech tagging, lemmatization, stemming and candidate filtration. A predefined list of stop words is used to identify stop words. There are several stop word lists available for the English language. For the sake of a fair evaluation and comparison, we selected the stopword list[2] used by the most recent approach by Duari and Bhatnagar [36]. Additionally, any words with less than three characters are considered stop words and are removed from the text.

CoCoNoW also introduces the Minimal Occurrence Frequency (MOF) which is inspired by Average Occurrence Frequency (AOF) [17]. MOF can be represented as follows:

$$\mathrm{MOF}(D, \beta) = \beta \frac{\sum\limits_{t \in D} \mathrm{freq}(t)}{|D|} \tag{3.1}$$

where β is a parameter, $|D|$ is the number of terms in the document D and $\mathrm{freq}(t)$ is the frequency of term t. The MOF supports some variation with the parameter β; a higher β means more words get removed, whereas a lower β means fewer words get removed. This allows customizing the CoCoNoW to the document length: Longer Documents contain more words, therefore, having a higher frequency of terms. Parameter optimization techniques on various datasets suggest that the best values for β are about 0.5 for short documents e.g. only analyzing abstracts of papers rather than the entire text; and 0.8 for longer documents such as entire papers.

3.2.1.2 *Graph Building*

CoCoNoW is a graph-based approach, it represents the text as a graph. We performed experiments with various window sizes for

[2] http://www.lextek.com/manuals/onix/stopwords2.html, accessed June-2022

CoCoNoW, including different numbers of consecutive sentences for the dynamic window size employed by CAG [36]. The performance dropped when more than two consecutive sentences were considered in one window. Therefore, a dynamic window size of two consecutive sentences was adopted for CoCoNoW. This means that an edge is added between any two terms if they occur within two consecutive sentences.

3.2.1.3 *Weight Assignments*

CoCoNoW is based on the Keyword Extraction using Collective Node Weight (KECNW) model developed by Biswas et al. [17]. The general idea is to assign weights to the nodes and edges that incorporate many different features, such as frequency, centrality, position, and weight of the neighborhood.

EDGE WEIGHTS The weight of an edge typically depends on the relationship it represents, in our case this relationship is term co-occurrence. Hence, the weight assigned to the edges is the normalized term co-occurrence $w(e)$, which is computed as follows:

$$w(e) = \frac{\text{coocc}(t_u, t_v)}{\text{maxCoocc}} \tag{3.2}$$

where the weight $w(e)$ of an edge $e = \{u, v\}$ is obtained by dividing the number of times the corresponding terms t_u and t_v co-occur in a sentence ($\text{coocc}(t_u, t_v)$) by the maximum number of times any two terms co-occur in a sentence ($maxCoocc$). This is essentially a normalization of the term co-occurrence.

NODE WEIGHTS The final node weight is a summation of four different features. Two of these features, namely *distance to most central node* and *term frequency* are also used by [17]. In addition, we employed *positional weight* [41] and the newly introduced *summary bonus*. All of these features are explained as follows:

Distance to most central node: Let c be the node with the highest degree. This node is considered the most central node in the graph. Then assign the inverse distance $D_C(v)$ to this node as the weight for all nodes:

$$D_C(v) = \frac{1}{d(c, v) + 1} \tag{3.3}$$

where $d(c, v)$ is the distance between node v and the most central node c.

Term Frequency: The number of times a term occurs in the document divided by the total number of terms in the document:

$$\text{TF}(t) = \frac{\text{freq}(t)}{|D|} \tag{3.4}$$

where freq(t) is the frequency of term t and |D| is the total number of terms in the document D.

Summary Bonus: Words occurring in summaries of documents, e.g. abstracts of scientific articles, are likely to have a higher importance than words that only occur in rest of the document:

$$SB(t) = \begin{cases} 0 & \text{if t does not occur in the summary} \\ 1 & \text{if t occurs in the summary} \end{cases} \tag{3.5}$$

where SB(t) is the summary bonus for term t. If there is no such summary, the summary bonus is set to 0.

Positional Weight: As proposed by Florescu and Caragea [41], words appearing in the beginning of the document have a higher chance of being important. The positional weight PW(t) is based on this idea and is computed as follows:

$$PW(t) = \sum_{j}^{freq(t)} \frac{1}{p_j} \tag{3.6}$$

where freq(t) is the number of times term t occurs in the document and p_j is the position of the j^{th} occurrence in the text.

FINAL WEIGHT COMPUTATION FOR COCONOW: The final node weight W uses all these features described above and combines them as follows:

$$W(v) = SB(t_v) + D_C(v) + PW(t_v) + TF(t_v) \tag{3.7}$$

where t_v is the term corresponding to node v, SB is the summary bonus, D_C is the distance to the most central node, PW is the positional weight and TF is the term frequency. All individual summands have been normalized in the following way:

$$norm(x) = \frac{x - minVal}{maxVal - minVal} \tag{3.8}$$

where x is a feature for an individual node, minVal is the smallest value of this feature and maxVal is the highest value of this feature. With this normalization, each summand in equation 3.7 lies in the interval $[0, 1]$. Thus, all summands are considered to be equally important.

3.2.1.4 *Node and Edge Rank (NER)*

Both the assigned node and edge weights are then used to perform Node and Edge Rank (NER) [12]. This is a variation of the famous PageRank [95] and is recursively computed as given below:

$$\mathrm{NER}(v) = (1-d)W(v) + dW(v) \sum_{e=(u,v)} \frac{w(e)}{\sum\limits_{e'=(u,w)} w(e')} \mathrm{NER}(u) \quad (3.9)$$

where d is the damping factor, which regulates the probability of jumping from one node to the next one [17]. The value for d is typically set to 0.85. $W(v)$ is the weight of node v as computed in equation 3.7. $w(e)$ denotes the edge weight of edge e, $\sum_{e'=(u,w)} w(e')$ denotes the summation over all weights of incident edges of an adjacent node u of v and $\mathrm{NER}(u)$ is the Node and Edge Rank of node u.

This recursion stops as soon as the absolute change in the NER value is less than the given threshold of 0.0001. Alternatively, the execution ends as soon as a total of 100 iterations are performed. However, it is just a precaution, as the approach usually converges in about 8 iterations. Mihalcea et al. [85] report that the approach needed about 20 to 30 steps to converge for their dataset. All nodes are then ranked according to their NER. Nodes with high values are more likely to be keywords. Each node corresponds to exactly one term in the document, so the result is a priority list of terms that are considered keywords.

3.2.2 *Stage 2: Topic Modeling*

In this section, we will discuss the second stage of our approach. The topic modeling task is increasingly popular on social web data [4, 10, 91, 114], where the topics of interest are unknown beforehand. However, this is not the case for the task in hand, i.e. clustering publications based on their topics. All publications share a common topic, for example, all International Conference on Document Analysis and Recognition (ICDAR) papers have *Document Analysis* as a common topic. Our proposed approach takes advantage of the common topic by incorporating an ontology. In this work, an ontology is used to define the possible topics where the detected keywords of each publication are subsequently mapped onto the topics defined by the ontology. For this task, we processed ICDAR publications from 1993 to 2017. The reason for selecting ICDAR publications for this task is that we already had the citation data available for these publications which will eventually be helpful during the evaluation of this task.

3.2.2.1 *Topic Hierarchy Generation*

All ICDAR publications fall under the category of *Document Analysis*. The first step is to find a suitable ontology for the ICDAR publications. For this purpose, the CSO 3.1[3] [107] was employed. This ontology

[3] https://cso.kmi.open.ac.uk, accessed Dec-2019

was built using the Klink-2 approach [93] on the Rexplore dataset [94] which contains about 16 million publications from different research areas in the field of computer science. These research areas are represented as the entities in the ontology. The reason for using this ontology rather than other manually crafted taxonomies is that it was extracted from publications with the latest topics that occur in publications. Furthermore, Salatino et al. [108] used this ontology already for the same task. They proposed an approach for the classification of research topics and used the CSO as a set of available classes. Their approach was based on bi-grams and tri-grams and computes the similarity of these to the nodes in the ontology by leveraging word embeddings from word2vec [86].

3.2.2.2 *Computer Science Ontology*

The CSO 3.1 contains 23, 800 nodes and 162, 121 edges. The different relations between these nodes are based on the Simple Knowledge Organization System[4] and include eight different types of relations.

3.2.2.3 *Hierarchy Generation*

For this task, we processed ICDAR publications from 1993 to 2017. Therefore, in line with the work of Breaux and Reed [21], the node *Document Analysis* is considered the root node for the ICDAR conference. This will be the root of the resulting hierarchy. Next, nodes are added to this hierarchy depending on their relations in the ontology. All nodes with the relation *superTopicOf* are added as children to the root. This continues recursively until there are no more nodes to add. Afterwards, three relation types *sameAs*, *relatedEquivalent* and *preferentialEquivalent* are used to merge nodes. The edges with these relations between terms describe the same concept, e.g. *optical character recognition* and *OCR*. One topic is selected as the main topic while all merged topics are added in the synonym attribute of that node. Note that all of these phrases are synonyms of essentially the same concept. The extracted keywords are later on matched against these sets of synonyms. Additionally, very abstract topics such as *information retrieval* were removed as they are very abstract and could potentially be a super-topic of most of the topics in the hierarchy thus making the hierarchy unnecessarily large and complicated. Lastly, to mitigate the missing topics of specialized topics like *Japanese Character Recognition*, we explored the official topics of interest for the ICDAR community[5]. An examination of the hierarchy revealed missing specialized topics like the only script dependent topic available in the hierarchy was *Chinese Characters*, so other scripts such as Greek, Japanese and Arabic were added as siblings of this node. We also created a default

[4] https://www.w3.org/2004/02/skos, accessed June-2022
[5] https://icdar2019.org/call-for-papers, accessed June-2022

node labeled *miscellaneous* for all those specialized papers which can not be assigned to any of the available topics.

Eventually, the final topic hierarchy consists of 123 nodes and has 5 levels. The topics closer to the root are more abstract topics while the topics further away from the root node represent more specialized topics.

3.2.2.4 *Topic Assignment*

Topics are assigned to papers by using two features of each paper: The title of the publication and the top 15 extracted keywords. The value of k=15 was chosen after manually inspecting the returned keywords; fewer keywords mean that some essential keywords are ignored, whereas a higher value means there are more unnecessary keywords that might lead to a wrong classification.

In order to assign a paper to a topic, we initialize the matching score with 0. The topics are represented as a set of synonyms, these are compared with the titles and keywords of the paper. If a synonym is a substring of the title, a constant of 200 is added to the score. The assumption is that if the title of a paper contains the name of a topic, then it is more likely to be a good candidate for that topic. Next, if all unigrams from a synonym are returned as keywords, the term frequency of all these unigrams is added to the matching score. By using a matching like this, different synonyms will have a different impact on the overall score depending on how often the individual words occurred in the text. To perform matching we used the Levenshtein distance [69] with a threshold of 1. This is the case to accommodate the potential plural terms. The constant bonus of 200 for a matching title comes from assessing the average document frequency of the terms. Most of the synonyms consist of two unigrams, so the document frequencies of two words are added to the score in case of a match. This is usually less than 200 - so the matching of the title is deemed more important.

Publications are assigned to 2.65 topics on average with a standard deviation of 1.74. However, the values of the assignments differ greatly between publications. The assignment score depends on the term frequency, which itself depends on the individual writing style of the authors. For this reason, the different matching scores are normalized: For each paper, we find the highest matching score, then we divide all matching scores by this highest value. This normalization means that every paper will have one topic that has a matching score of 1 - and the scores of other assigned topics will lie in the interval $(0, 1]$. This accounts for different term frequencies and thus, also the different writing styles.

Table 3.1: Distribution details of the datasets

Dataset	|D|	L	Avg / SD	Dataset Description
Hulth2003 [57]	1500	129	19.5 / 9.98	Abstracts
NLM500 [7]	500	4854	23.8 / 8.19	Full papers
SemEval2010 [64]	244	8085	25.5 / 6.96	Full papers

3.3 EVALUATION

In this section, we will discuss the experimental setup and the evaluation of our system where we firstly discuss the evaluation of our first stage CoCoNoW for keyword detection. The results from CoCoNoW are compared with various state-of-the-art approaches on three different datasets: Hulth2003 [57], SemEval2010 [64] and NLM500 [7]. Afterwards, we will discuss the evaluation results of the second stage for topic modeling as well.

3.3.1 *Experimental Setup*

Keyword detection approaches usually return a ranking of individual keywords. Hence, the evaluation is based on individual keywords. For the evaluation of these rankings, a parameter k is introduced where only the top k keywords of the rankings are considered. This is a standard procedure to evaluate performance [36, 64, 74, 79, 121].

However, as the gold-standard keywords lists contain key phrases, these lists undergo a few preprocessing steps. Firstly, the words are lemmatized and stemmed, then a set of strings called the evaluation set is created. It contains all unigrams. All keywords occur only once in the set, and the preprocessing steps allow the matching of similar words with different inflections. The top k returned keywords are compared with this evaluation set.

Note that the evaluation set can still contain words that do not occur in the original document, which is why an F-Measure of 100% is infeasible. For example, the highest possible F-Measure for the SemEval2010 dataset is only 81% because 19% of the gold standard keywords do not appear in the corresponding text [64].

Table 3.1 gives an overview of performance on different datasets. These datasets were chosen because they cover different document lengths, ranging from about 130 to over 8000 words and belong to different domains: biomedicine, information technology, and engineering.

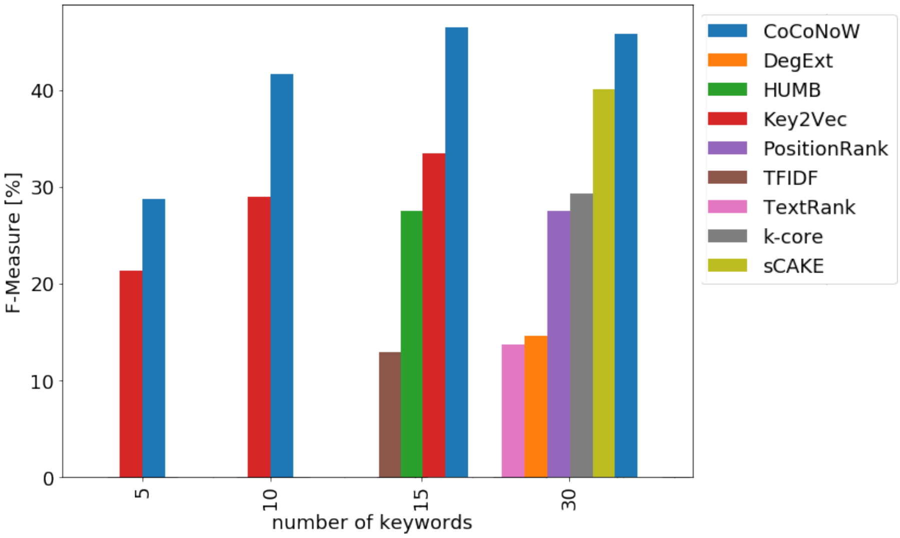


Figure 3.2: Evaluation of CoCoNoW on the SemEval2010 [64] dataset.

3.3.2 *Performance Evaluation Stage 1: CoCoNoW*

The performance of keyword detection approaches is assessed by matching the top k returned keywords with the set of gold standard keywords. The choice of k influences the performance of all keyword detection approaches as the returned ranking of keywords differs between these approaches. Table 3.2 shows the performance in terms of Precision, Recall and F-Measure of several approaches for different values of k. Fig. 3.2 and Fig. 3.3 compare the performance of our approach with other approaches on the **SemEval2010** [64] dataset and the **Hulth2003** [57] dataset respectively.

By looking at the results of all trials, we can make the following observations:

- CoCoNoW always has the highest F-measure
- CoCoNoW always has the highest Precision
- In the majority of the cases, CoCoNoW also has the highest Recall

For the **Hulth2003** [57] dataset, CoCoNoW achieved the highest F-measure of 57.2% which is about 6.8% more than the previous state-of-the-art. On the **SemEval2010** [64] dataset, CoCoNoW achieved the highest F-measure of 46.8% which is about 6.2% more than the previous state-of-the-art. Lastly, on the **NLM500** [7] dataset, CoCoNoW achieved the highest F-measure of 29.5% which is about 1.2% better than previous best performing approach. For $k = 5$ CoCoNoW achieved the same Recall as the Key2Vec approach by Mahata et al. [79], however, the Precision was 15.2% higher. For $k = 10$, the approach by Wang et al. [121] had a Recall of 52.8%, whereas CoCo-

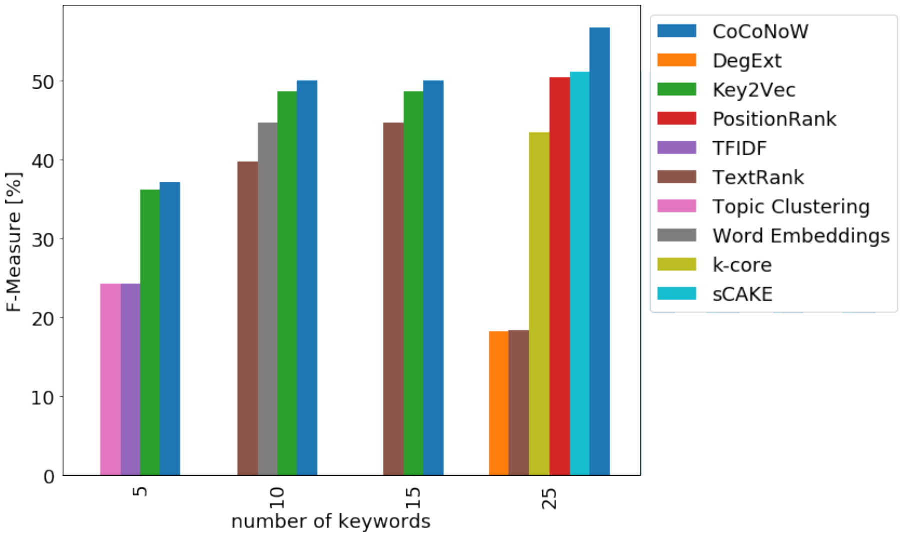


Figure 3.3: Evaluation of CoCoNoW on the Hulth2003 [57] dataset.

Table 3.2: Performance comparison of CoCoNoW with several other approaches on the different datasets.
* Results reported by Duari and Bhatnagar [36]

Approach	k	Hulth2003 [57]			SemEval2010 [64]			NLM500 [7]		
		P[%]	R[%]	F1[%]	P[%]	R[%]	F1[%]	P[%]	R[%]	F1[%]
TF-IDF [74]	5	33.3	17.3	24.2	-	-	-	-	-	-
Topic Clustering [74]		35.4	18.3	24.3	-	-	-	-	-	-
Key2Vec [79]		68.8	**25.7**	36.2	41.0	14.4	21.3	-	-	-
CoCoNoW		**84.0**	**25.7**	**37.3**	**84.1**	**17.5**	**28.7**	48.8	11.4	17.9
TextRank [85]	10	45.4	47.1	39.8	-	-	-	-	-	-
Word Embeddings [121]		38.7	**52.8**	44.7	-	-	-	-	-	-
Key2Vec [79]		57.6	42.0	48.6	35.3	24.7	29.0	-	-	-
CoCoNoW		**73.3**	41.9	**50.0**	**72.3**	**29.8**	**41.6**	43.3	19.8	26.3
supervised approach [57]	16	25.2	51.7	33.9	-	-	-	-	-	-
TextRank [85]	14	31.2	43.1	36.2	-	-	-	-	-	-
TF-IDF [64]	15	-	-	-	11.6	14.5	12.9	-	-	-
HUMB [77]	15	-	-	-	27.2	27.8	27.5	-	-	-
Key2Vec [79]	15	55.9	50.0	52.9	34.4	32.5	33.4	-	-	-
CoCoNoW	15	**63.5**	**52.9**	**54.2**	**62.2**	**39.2**	**46.5**	37.11	25.2	29.0
TextRank [85]	25	-	-	18.4*	-	-	-	-	-	-
DegExt [73]		-	-	18.2*	-	-	-	-	-	-
k-core [104]		-	-	43.4*	-	-	-	-	-	-
PositionRank [41]		45.7*	64.5*	50.4*	-	-	-	-	-	-
sCAKE [36]		45.4	**66.8**	51.1	-	-	-	-	-	-
CoCoNoW		**54.8**	66.2	**56.8**	47.3	47.8	46.8	29.3	32.6	29.9
TextRank [85]	30	-	-	-	-	-	13.7*	-	-	10.7*
DegExt [73]		-	-	-	-	-	14.6*	-	-	10.9*
k-core [104]		-	-	-	-	-	29.3*	-	-	20.2*
PositionRank [41]		-	-	-	25.3*	31.3*	27.5*	19.7*	26.6*	21.9*
sCAKE [36]		-	-	-	35.8	47.4	40.1	24.5	35.0	28.3
CoCoNoW		52.5	70.1	57.2	**42.6**	**51.5**	**45.8**	**26.7**	**35.3**	**29.5**

NoW only achieved 41.9%; this results in a difference of 10.9%. However, the Precision of CoCoNoW is almost twice as high, i.e. 73.3%

as compared to 38.7%. Lastly, for $k = 25$, the sCAKE algorithm by Duari and Bhatnagar [36] has a higher Recall (0.6% more), but also a lower Precision (9.4% less). All in all, CoCoNoW extracted the most keywords successfully and outperformed other state-of-the-art approaches. Note the consistently high Precision values of CoCoNoW: There is a low number of false positives (i.e. words wrongly marked as keywords), which is crucial in the next stage of clustering publications with respect to their respective topics.

3.3.3 *Performance Evaluation Stage 2: Topic Modeling*

This section discusses the evaluations of ontology-based topic modeling. There was no ground truth available for the ICDAR publications, consequently making the evaluation of topic modeling a challenging task. Nevertheless, we employed two different approaches for evaluation: manual inspection and citation count. Details of both evaluations are as follows:

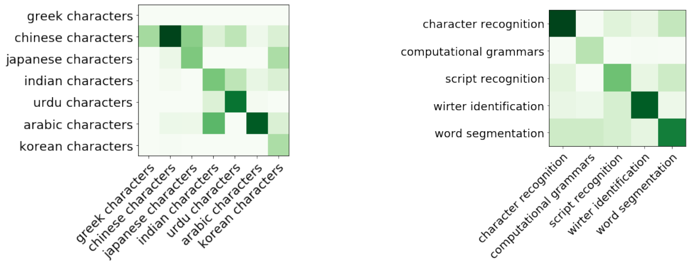
3.3.3.1 *Manual Inspection*

The proposed method for topic assignment comes with labels for the topics, so manually inspecting the papers assigned to a topic is rather convenient. This is done by going through the titles of all papers assigned to a topic and judging whether the assignment makes sense.

For specialized topics i.e. the ones far from the root, the method worked very well, as it is easy to identify papers that do not belong to a topic. Manual inspection showed that there are very few false positives, i.e. publications assigned to an irrelevant topic. This is because of the high Precision of the CoCoNoW algorithm: The low number of false positives in the extracted keywords increases the quality of the topic assignment. The closer a topic is to the root i.e. a more generic topic, the more difficult it is to assess whether a paper should be assigned to it: Often, it is not possible to decide whether a paper can be assigned to a general node such as *neural networks* by just reading the title. Hence, this method does not give meaningful results for more general topics. Furthermore, this method was only able to identify false positives. It is difficult to identify false negatives with this method, i.e. publications not assigned to relevant topics.

3.3.3.2 *Evaluation by Citation Count*

The manual inspection indicated that the topic assignment works well, but that was just a qualitative evaluation. So a second evaluation is performed. It is based on the following assumption: *Papers dealing with a topic cite other papers from the same topic more often than papers dealing with different topics.* We believe that this is a sensible assumption, as papers often compare their results with previous approaches



(a) *Character Recognition*

(b) *Pattern Recognition*

Figure 3.4: Citation count for different super topics.

that tackled similar problems. So instead of evaluating the topic assignment directly, it is indirectly evaluated be counting the number of citations between the papers assigned to the topics. However, a paper can be assigned to multiple overlapping topics (e.g. *machine learning* and *neural networks*). This makes it infeasible to compare topics at two different hierarchy levels using this method.

Nevertheless, this method is suited for evaluating the topic assignment of siblings, i.e. topics that have the same super topic. Figure 3.4 shows heatmaps for citations between child elements of a common supertopic in different levels in the hierarchy. The rows represent the number of citing papers, the columns the number of cited papers. A darker color in a cell represents more citations. Fig. 3.4a shows the script-dependent topics, which are rather specialized and all have the common supertopic *character recognition*. The dark diagonal values clearly indicate that the number of intra-topic citations are higher than inter-topic citations.

Figure 3.4b shows the subtopics of the node *pattern recognition*, which has a very high level of abstraction and is close to the root of the hierarchy. The cells in the diagonal are clearly the darkest ones again, i.e., there are more citations within the same topic than between different topics. This is a recurring pattern across the entire hierarchy. So in general, results from these evaluations indicate that the topic assignment works reliably therefore suggesting that our assumption is correct.

Part III

DATA ANALYSIS

4 SENTIMENT & INTENT ANALYSIS

Scientific publications play an important role in the development of a community. An exponential increase in scientific literature has posed the challenge of evaluating the impact of a publication in a given scientific community. Citations majorly contribute towards the eminence of an author as well as the impact of their publications on society. However, citation count serves as a quantitative metric and therefore does not provide qualitative insights into the citations. To get a qualitative insight, the sentiment of a given citation needs to be identified, where the citation sentiment refers to the opinion of the citing author about the cited literature.

This chapter emphasizes that using a qualitative metric by taking into account different aspects of the citation leads to a much more sophisticated representation of the importance of a citation. Therefore, the sentiment and intent are used exemplary as two meaningful features that can enhance the effectiveness of the currently employed quality assessment approaches. The quality of a research artifact depends on the content and the results of a publication and its acceptance in the research community. In short, to create a good metric, it is important to cover additional aspects independent of the number of citations.

Sentiment classification provides us contextual insight into each literature citation. Sentiment classification is commonly applied to different domains [9, 40, 71, 82, 125] i.e., movie reviews, product reviews, citations, etc. where a given text string is classified based on its inherent sentiment. Thus, it is possible to classify sentiments as either subjective & objective or a more fine-grained classification into positive, neutral, and negative depending on the domain and instances. However, sentiment classification can also induce subjectivity to the opinion.

Sentiment classification provides us with a deeper qualitative insight into a given literature citation. However, to get even deeper insights and to evade the likelihood of subjectivity, intent could also be identified. The intent of a literature citation refers to the purpose of citing the existing literature. An author can cite a published manuscript for several reasons, i.e., describing related works, employing, extending, or comparing existing approaches, and contradicting the claims from previous literature. Intent classification plays a crucial role in validating the predicted sentiment of a given citation. The positioning of the citation plays an important role in identifying the sentiment. For instance, citations usually found in the evaluation and discussion section are more likely to be negative, as the citing authors usually compare the results of their approach in evaluation to prove the superiority of their approach.

Despite the recently published approaches e.g., Beltagy et al. [13] there is still a lack of methods and datasets used for scientific citation analysis. This lack of data originates from the effort mandatory to annotate scientific citations. Furthermore, most sentiment analyses cover domains in which the data is highly subjective, and the annotation can be automated. Besides, there is no common definition of intention used to classify publications properly. In this work, we cleaned a publicly available dataset for the task of citation sentiment analysis and benchmarked the performance of several models ranging from simple Convolutional Neural Network (CNN) to more sophisticated transformer networks for sentiment and intent classification. By doing so, we achieved a new state-of-the-art for both sentiment and intent classification. The contributions of this chapter are as follows:

- We propose one solution for both tasks in hand i.e. sentiment and intent classification. The proposed model can be separately trained for both tasks.
- We removed the discrepancies and the redundancies present in the previous version of an existing dataset and publicly released a clean and reliable dataset for citation sentiment analysis[1]
- We conducted performance benchmarking of a set of models ranging from simple CNN-based models to sophisticated transformer networks and achieved state-of-the-art performance for both sentiment and intent classification.
- We performed an evaluation of out-domain data usage during training
- We also evaluated different scheduling methods
- Lastly, this chapter also proposes an end-to-end sentiment and intent citation classification multitask model

4.1 LITERATURE REVIEW

In this section, we discuss the existing literature for sentiment and intent classification. We also highlight the key aspects of each existing approach.

[1] https://github.com/DominiqueMercier/ImpactCite

4.1.1 *Sentiment Classification*

Sentiment classification is a popular task and due to its wide range of applications, there exist numerous publications which address this problem. Tang et al. [117] proposed sentiment-specific word embeddings for performing sentiment classification of tweets and highlighted that the use of highly specialized word embeddings can improve performance for sentiment classification. Thongtan et al. [118] employed document embeddings trained with cosine similarity to perform sentiment classification on a movie review dataset. Cliche [30] proposed a sentiment classifier for tweets consisting of an ensemble of CNN and LSTM models trained and finetuned on a large corpus of unlabeled data.

With the popularity of transformer networks, Bidirectional Encoder Representations from Transformers (BERT)[35] became a famous choice among the community for a range of Natural Language Processing (NLP) tasks. The BERT model was trained on a large volume of unlabeled data. Hence, recent literature in the sentiment analysis domain employed the BERT model to improve the performance of the task at hand. In [89, 126, 131], the authors take advantage of transfer learning to adapt the pre-trained BERT model for sentiment classification and further boost the performance by complementing it with preprocessing, attention modules, structural features, etc.

The literature discussed so far dealt with sentiment classification in tweets or movie reviews. On the other hand, citation sentiment classification is quite different from movie or product review sentiment classification, as the text in scientific publications is formal. Esuli and Sebastiani [39] defined that the sentiment classification is analogous to opinion mining and subjectivity mining. They further discussed that the personal preferences and writing style of an author can induce subjectivity in the citations as an author can deliberately make a citation sound positive or negative. Athar [8] performed different experiments using sets of various features like science lexicon, contextual polarity, dependencies, negation, sentence splitting, and word-level features to identify an optimal set of features for sentiment classification in scientific publications. Xu et al. [127] performed sentiment analysis of citations in clinical trial papers by using textual features like n-grams, sentiment lexicon, and structure information. Sentiment classification is significantly important in the domain of scientific citation analysis due to the scarcity of scientific datasets suitable for citation sentiment classification and the shallow definition of sentiment for this domain. Finding a sentiment in a text that is written to be analytical and objective is substantially different from doing so in highly subjective text pieces like Twitter data.

4.1.2 *Intent Classification*

The basic concepts of intent classification are the same as sentiment classification. However, contrary to the sentiment classification, the definition of the citation intent classification is much sharper and the label acquisition is strongly related to the sections of a paper where it appears. Usually, the section title provides a good understanding of the intent of the citation. However, compound section titles in scientific work can prove to be challenging for identifying the intent. Cohan et al. [31] performed citation intent analysis by employing bi-directional LSTM with attention mechanism and consolidating it with ELMo vectors and structural scaffolds like citation worthiness and section title.

Beltagy et al. [13] proposed *SciBERT*, which is a variation of BERT optimized for scientific publications and trained on 1.14 Million scientific publications containing 3.17 Billion tokens from biomedical and computer science domains. SciBERT was applied to a group of NLP tasks including text classification to sections. Mercier et al. [83] employed a fusion of Support Vector Machine (SVM) and perceptron-based classifier to classify the intent of the citations. They used a set of textual features consisting of the type & length of tokens, capitalization, adjectives, hypernyms, and synonyms. Similarly, Abu-Jabra et al. [2] also employed SVM to perform the intent classification of citations. They suggested that lexical and structural features play a crucial role in identifying the intent of a given citation.

4.1.3 *Out-Domain Data Utilization*

Su et al. [116] presented in their work to study the impact of out-domain data for question answering. They investigated different training schedules and their impact on accuracy. The main focus of their work was a better generalization. Another work that conducted experiments related to the robust training using in-domain and out-domain data was proposed by Li et al. [70]. Their proposed method provides the capabilities to learn domain-specific and general data in conjunction to overcome the convergence towards domain-specific properties. Sajjad et al [106] proposed an approach that first learns of different out-domain data and finally fine-tunes on in-domain data to achieve the optimal results. This approach intuitively utilizes the data of the different domains and therefore has a much larger training corpus for a better generalization.

Khayrallah et al. [63] addressed the amount of out-domain vocabulary. Their findings showed that with the use of out-domain data and a continuous adaption of the domain, the number of words not included in the vocabulary can be reduced efficiently. For this purpose, they used an out-domain model and trained it with a modified training objective continuously on the in-domain data. Furthermore, Mrkšić et al. [88] showed that using the out-domain data can yield significant improvements for very small datasets. And therefore makes it possible to train models using these sets when it is not possible to do that without the use of out-domain data.

4.2 DATASETS

This chapter focuses on the task of sentiment and intent analysis. Therefore, we selected a range of datasets suitable for sentiment classification and also for intent classification. We identified some inconsistencies in the citation sentiment dataset, which were later addressed and a clean version of that dataset was released along with this work. However, despite the dataset limitation we decided to stick with the sentiment dataset and improve its quality to propose a cleaned version usable to perform citation sentiment analysis using deep neural network models.

4.2.1 *Sentiment Datasets*

For the task of Sentiment classification, we employed various datasets for our experiments. Our target domain is the scientific literature and there exist very limited publicly available datasets for citation sentiment analysis. Therefore, we selected some out-domain datasets to overcome the data scarcity. Following are the datasets selected for the sentiment classification task:

1. Movie reviews
2. Product reviews
3. Twitter data
4. Scientific data

To standardize the labels of selected datasets, a preprocessing step was essential. For experiments evaluating out-domain knowledge transfer and sequential training, we preprocessed the selected datasets for binary sentiment classification tasks i.e. positive and negative. It enabled us to train and test models across different datasets. To do so, we excluded the neutral class and grouped different labels if the datasets had multiple classes that correspond to the positive or negative label e.g. ’good’ and ’very good’ or 4 out of 5 and 5 out of 5 stars. However, we used all three classes i.e. positive, negative, and neutral for the multi-task experiments. The details of the selected sentiment datasets are as follows:

4.2.1.1 *Movie Reviews:*

From the domain of movie reviews, we decided to use three popular datasets that quantified both positive and negative reviews in the form of a numerical score. The IMDB [78] dataset contains about $25,000$ training and $25,000$ test instances of highly polar reviews. It is the largest dataset by volume in the selected datasets. The second dataset we used in our experiments is the Cornell movie review data [96]. It is a considerably small dataset as compared to IMDB. However, it has an even distribution of $1,000$ samples for each of the positive and negative classes. The last dataset that we selected from movie reviews is the Stanford Sentiment Treebank dataset [115]. For this dataset, we had to discard the samples not related to either negative or positive classes. All three above-mentioned datasets are related to the same task and the same domain and therefore their underlying structure should be rather similar.

4.2.1.2 *Product Reviews:*

To include a dataset from a different domain than the Movie reviews, we selected the Amazon product review dataset [81]. This dataset consists of various product categories. Some categories in the Amazon data are closely related to the movie reviews such as Books, TV, and Movies. On the other hand, some categories are completely different from movie reviews such as Beauty, Electronic, and Video Games. For our experiments, we selected one category from Amazon data that was unrelated to the movie reviews. The chosen category was related to the instrument reviews. The product reviews were quantified in the form of $1-5$ stars. For our experiments, we converted the star ratings into positive and negative classes while skipping the neutral class. Product reviews with ratings with 4 and 5 stars were labeled as positive. On the other hand, product reviews with 1 and 2 stars were labeled as negative. However, product reviews with a star rating of 3 were skipped as they belonged to the neutral class and were not relevant for our experiments.

4.2.1.3 *Twitter Data:*

Sentiment analysis on Twitter data is a quite popular task. For this purpose, we selected a couple of Twitter datasets. Intuitively, we assume that the Twitter datasets are the most subjective ones in our selection as their language style differs significantly from the scientific and other domain datasets. The first dataset is related to airline reviews in form of tweets. The dataset was taken from Kaggle [2] and contains three classes i.e. positive, negative, and neutral. Similar to

[2] Twitter US Airline Sentiment: https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Table 4.1: Citation sentiment corpus [8]. Number of instances and class distribution.

	Classes		
	Positive	**Negative**	**Neutral**
Avg. Length	229.4	221.8	219.6
No. of samples	829	280	7627
Class dist.	9.49%	3.21%	87.30%

other out-domain datasets, we removed the neutral class. The same class elimination was performed for the second dataset Sentiment140 dataset [3]. This dataset was composed using 1.6 Million general tweets collected from Twitter along with their sentiment.

4.2.1.4 *Scientific Data:*

CSC: A Citation Sentiment Corpus: When it comes to the task of citation sentiment classification using publicly available high-quality datasets there is a severe lack of data availability. Although there exist datasets for scientific papers e.g. the dataset proposed by Xu et al. [127] or the sentiment citation corpus proposed by Athar [8] these are either not publicly available or have quality issues. Precisely, this problem originates due to the difficulty in data acquisition and labeling of scientific text as it can not be automated. Conversely, it is straightforward to acquire Twitter or movie review data and label it. Due to the lack of alternate solutions, we had to stick to the dataset proposed by Athar [8] although this dataset has a very unbalanced class distribution as shown in Table 4.1. Fig 4.1 shows the distribution of samples among different classes. In the following sections, we refer to this dataset as Citation Sentiment Corpus (CSC).

The CSC dataset consists of three classes Positive, Negative and Neutral. Where each class label represents the opinion of the citing author about the cited literature. Fig 4.2 shows the variation in the length of samples and their distribution among different classes. Generally, a citation contains multiple sentences resulting in an additional context that can be utilized for judging its sentiment. Extracting only the sentence containing the citation would result in a potential information loss as the sentiment can be included in a follow-up or previous sentence. Therefore, we decided to keep the instances as they are providing us instances of multiple sentences to assure that the content relation can be learned correctly.

CSC-Clean: A Cleaned Citation Sentiment Corpus: During the experimentation on the CSC dataset, we identified several discrepancies concerning duplicated instances, wrong data splits, and samples with

[3] Sentiment140: https://www.kaggle.com/kazanova/sentiment140

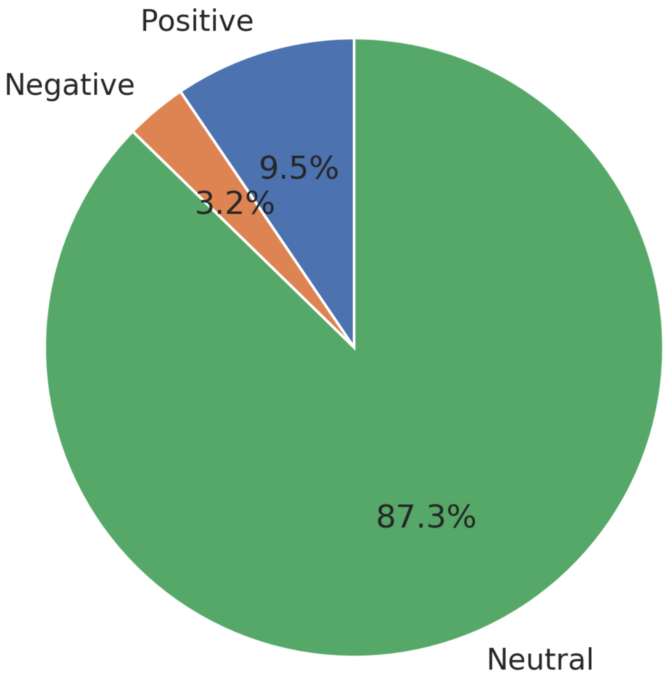


Figure 4.1: Citation sentiment corpus class distribution.

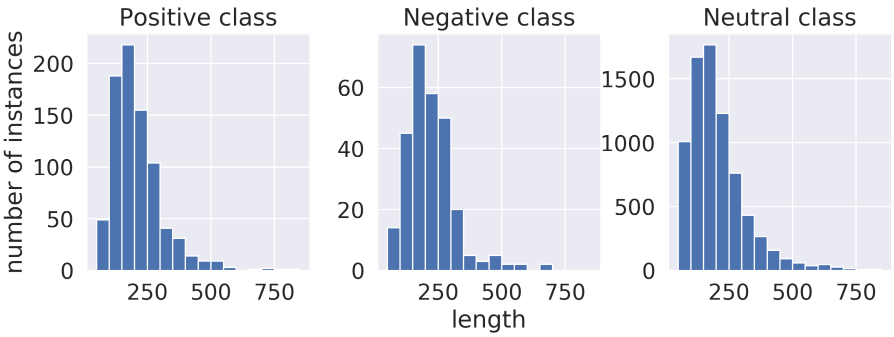


Figure 4.2: Citation sentiment corpus. Sample length class-wise.

impressively bad quality concerning their label consistency. Therefore, it was not possible to compare our approach with the existing results published for the citation sentiment corpus, and we decided to clean the dataset to create an improved version of the dataset with better quality covering the same corpus. For this purpose, we applied the following two steps for dataset cleansing:

1. Removing duplicate samples with different labels
2. Removing duplicate samples with the same labels

During dataset cleansing, we removed 756 instances as shown in Table 4.2. The removed instances were either identical duplicates of existing instances or provided different labels for the same text. In the case of samples with inconsistent labels, we removed all appearances

Table 4.2: Comparison of citation sentiment corpus and clean citation sentiment dataset.

	Classes		
	Positive	**Negative**	**Neutral**
Citation sentiment corpus	829	280	7627
Clean citation sentiment dataset	728	253	6999
Removed instances	101	27	628

Table 4.3: Comparison of citation sentiment corpus (CSC) and citation sentiment clean (CSC-C) dataset. Taken from [1].

Classes	**CSC**	**CSC-Clean**	**CSC-Clean Dist.**	**Removed [%]**
Positive	829	728	9.12%	101 (12.18)
Neutral	280	253	87.71%	27 (9.64)
Negative	7,627	6,999	3.17%	629 (8.25)

as a manual selection of a specific instance would induce a bias. Although this reduces the number of available instances, however, it is the most appropriate solution to exclude possible subjectivity. When it comes to the decision of which instances label is correct for the evaluation it is not suitable to keep both instances. We propose the dataset without any duplicates or inconsistent labels enabling us to produce fair and meaningful results using cross-validation to overcome the limited amount of instances for the minority classes. In this chapter, we will refer to this dataset as CSC-Clean. The cleaned dataset is publicly available on the following link: https://github.com/DominiqueMercier/ImpactCite.

In Table 4.4 we show the statistics of each sentiment dataset after pre-processing them to exclude the neutral class and existing duplicates. These statistics include the number of samples used to train, validate, and test our models. In addition, the table also shows the dataset distribution highlighting that datasets such as the Instruments, US Airline, and CSC-Clean is heavily biased towards one of the two classes. Another characteristic is that the collected datasets differ largely in their size. This resulted in the need to upsample or downsample the data for some experiments to make the results comparable.

4.2.2 *Intent Dataset*

From the scientific domain, we selected a dataset related to citation intent analysis called SciCite. The SciCite dataset proposed in [31] is a famous benchmark for citation intent classification. It was curated using medical and computer science publications and is publicly available. The size of this dataset is sufficient to train any deep learning model and the existing benchmarks emphasize the high quality of

Table 4.4: Comparison of all used datasets. Only including the positive and negative class. Neutral class for CSC-Clean was excluded in this table.

Domain	**Dataset**	**Train**	**Val**	**Test**	**+ve [%]**	**-ve [%]**
Movie Reviews	IMDB	19,923	4,981	24,678	50.19	49.81
	Cornell	6,823	1,706	2,133	50.0	50.0
	Stanford Sent.	6,911	872	1,819	51.64	4s8.36
Product Reviews	Instruments	6,068	1,507	1,897	95.07	4.93
Twitter Data	US Airline	7,243	1,811	2,264	19.81	80.19
	Sentiment140	10,161	2,541	3,176	49.94	50.06
Scientific Data	CSC-Clean	797	89	95	74.21	25.79

Table 4.5: SciCite [31]. Number of instances and class distribution. Taken from [1].

Classes	**Training**	**Validation**	**Test**	**Total**	**Percentage**
Result	1,109	123	259	1,491	13.53
Method	2,294	255	605	3,154	28.62
Background	4,840	538	997	6,375	57.85

the dataset. However, the dataset has an imbalanced sample distribution in which the vast majority of the samples are assigned to the 'Background' class. Another, important aspect of the dataset is the coarse-grained label process that was applied to create that dataset. According to the authors, the distribution follows the real-world distribution and the number of samples is large enough to sufficiently learn the concepts of each class. Detailed information about the dataset can be found in Table 4.5. We mainly employed SciCite along with the CSC-Clean dataset to demonstrate the capability of training a multi-task model, where tasks are different and yet from the same domain.

4.3 CONTRIBUTIONS

We divided this section into three parts. The first part discusses the proposed baseline approach for sentiment and intent analysis called ImpactCite [1]. The second part discusses the impact of training a model on out-domain data. And the third part covers a fusion approach to combine sentiment and intent analysis tasks. We further show that both methods rely on different aspects of the task and highlight their advantages.

4.3.1 *Citation Analysis Based on XLNet*

To tackle the problem of sentiment and intent analysis we propose ImpactCite, an XLNet-based approach. XLNet is a popular choice for

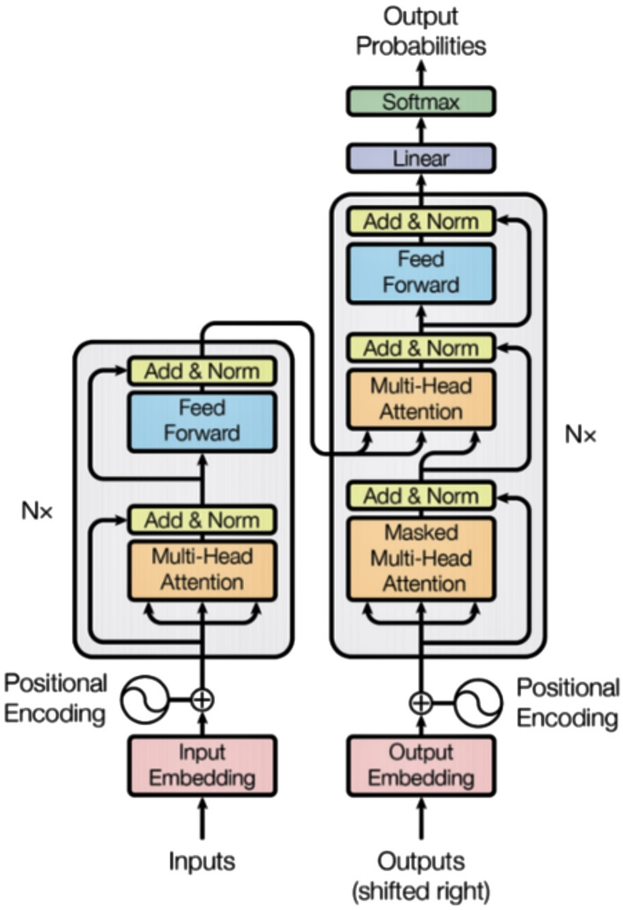


Figure 4.3: Transformer-XL architecture [34]. Each of the Multi-Head Attention layers is composed of multiple attention heads that apply a linear transformation and compute the attention.

several NLP-related tasks [129]. XLNet is an auto-regressive language model that contains bi-directional attention and is pre-trained on a large amount of data. The bi-directional attention makes it possible to understand relations within the sentences that can be drawn from left to right and vice versa. Due to the permutation generalization approach and the use of Transformer-XL [34] as the backbone model, XLNet can achieve excellent performance for language tasks involving long context. The Transformer XL architecture is shown in Fig 4.3. Especially, the capability to handle long context is important for the sentiment classification task as the sentiment of a citation can depend on the content of preceding or the proceeding sentences.

There are several variations of XLNet that differ slightly in the number of layers and units. For our experiments, we decided to use two XLNet-Large models. As our tasks cover a long context we decided to use the large version of XLNet. XLNet-Large consists of 24-layers, 1024 hidden units, and 16 heads. During our experiments, we rely on a pre-trained version of the model and fine-tune it according to the citation classification task. We start with a warm-up phase using a fixed learning rate followed by a slow learning rate decay to adjust the weights. This makes it possible to fine-tune the large model on a small dataset as the general language structure is already learned by the pre-trained model, and we only adjust the weights to the new domain and task.

In this work, we used one model for the sentiment classification and the second model for the intent classification. Separating these two tasks enables us to fine-tune the corresponding model to each task and achieve the best possible results for that task. This is especially beneficial for the intent as the amount of sentiment citation data is limited. However, the major drawback is that two separate models are required for this purpose and the sentiment does not benefit from the intent model, although both tasks are from the same domain.

4.3.2 *Overcoming Data Scarcity & Data Feeding Techniques*

In this section, we investigated the techniques to overcome the scarcity of data for certain domains. Particularly for sentiment analysis of scientific citations, there are not many datasets available. In this chapter, we propose that training on out-domain data and later finetuning on target domain results in better model performance, therefore, bridging the data scarcity gap. Additionally, we experimented with different data feeding methods to analyze their impact on the performance of the final model.

4.3.3 *Sentiment and Intent Fusion*

Lastly, in this section, we propose that although the citation sentiment and intent analysis are different tasks. However, we believe that the underlying text structure concerning the sentiment and intent task on scientific data is similar. Based on the cross-domain sentiment classification we show that the addition of data addressing the same task or the same domain can enhance the scientific sentiment classification. Ultimately, we train a single XLNet model on both the sentiment and intent datasets that performs the complete citation analysis and resolves the dataset size issues. The pipeline is visualized in Fig 4.4.

4.4 EXPERIMENTS AND ANALYSIS

In this section, we will discuss our experiments and their results. All experiments are classified into four sets. The first set discusses the performance benchmark of XLNet [129] for the task of intent classification. The second set discusses the experiments related to the performance benchmark of XLNet for sentiment classification. We performed these benchmarking experiments using several other models ranging from the baseline models i.e. CNN to highly sophisticated

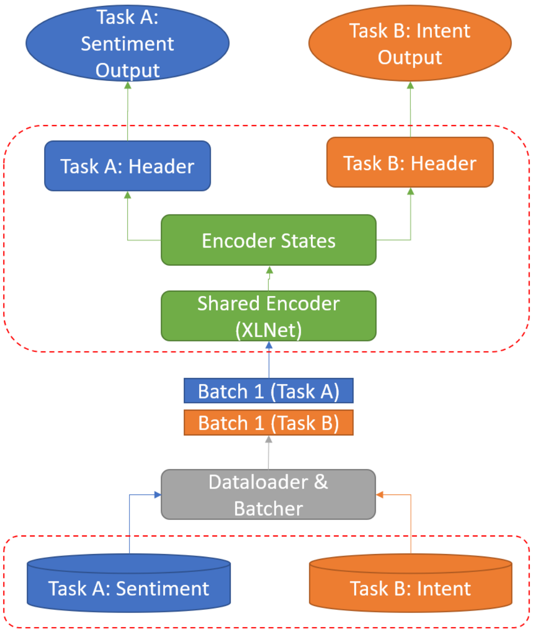


Figure 4.4: Mutli-Task setup combining sentiment and intent task. The same encoder is used for both tasks and a task specific head is trained.

language models i.e. BERT [35], A Lite Bidirectional Encoder Representations from Transformers (ALBERT) [67] and XLNet [129]. In the third set of experiments, we will discuss the experiments related to training on out-domain data and testing on several different domains dataset, which also includes finetuning on the target domain dataset. Additionally, a collection of experiments discussing the effects of different data feeding techniques are also discussed in this set of experiments.

Finally, we discuss experiments combining the sentiment and intent modality and serve a single model that processes both tasks. Doing so requires a deep understanding of multiple aspects such as domain dependency, model selection, and task relation. In our case, the first two aspects are covered by benchmarking and the out-domain evaluation. In addition, it has been shown that the tasks are related to each other [83].

4.4.1 *Intent Classification*

4.4.1.1 *Experiment1: Performance Benchmarking.*

To evaluate the performance of different model architectures on the intent classification task we decided to use the SciCite dataset [31]. We used the original train and test splits provided by the dataset and divided our models into two categories. The first category includes all baseline models. We explored different setups of CNNs, Long Short Term Memory (LSTM)s, and Recurrent Neural Network (RNN)s. These models were trained from scratch using the SciCite dataset. In addition, we also trained BERT [35], ALBERT [67] and ImpactCite (XLNet). The second category of models was pre-trained and is a member of the transformer-based solutions. These models were only fine-tuned on the SciCite dataset. Due to the high imbalance of data, we employed the micro-f1 and macro-f1 scores for performance comparison. Furthermore, initial experiments using the CNN, LSTM, and RNN approaches have shown their performance using pre-trained embeddings e.g. Global Vectors (GloVe)[4] did not improve compared to newly initialized embeddings. We emphasize that one of the reasons for this might be the domain discrepancy between the pre-trained embeddings and the scientific domain.

4.4.1.2 *Results and Discussion:*

Table 4.6 shows the performance benchmark results of different selected architectures for the intent classification task. It is evident from the results that both the LSTM and RNN are not able to compete with the CNN. A reason for the inferior performance of the RNN is the length of the sequences resulting in vanishing gradients for the RNN. The LSTM on the other hand suffers from the bi-directional influences between the sentences that are not completely covered by the architecture. We further explored different layer and filter sizes for baseline models. However, there is only an insignificant difference when tuning the parameters. Concerning the time consumption, the CNN shows superior performance over the other baseline approaches as it can compute things in parallel as compared to LSTMs and RNNs.

The second category presented in Table 4.6 shows the complex language models. We were able to achieve a new state-of-the-art performance using ImpactCite [1]. It significantly outperformed the other fine-tuned language models by up to 3.9% micro-f1 and 5.8% macro-f1 score. Especially, the increase in the minority classes has shown a significant difference of 10%. Summarizing the findings, we have demonstrated that ImpactCite (XLNet) was able to outperform the CNN by 8.71% and the language models by 3.9% macro-f1 score and significantly increased the performance of the minority class. This highlights the significantly better capabilities of the larger transformer-based model pre-trained on a different domain and later fine-tuned.

4.4.2 *Sentiment Classification*

In this section, we will discuss the experiments conducted for the task of scientific sentiment classification. There were two datasets used in these experiments namely Citation Sentiment Corpus (CSC) and our proposed clean version of the dataset called CSC-C.

[4] https://nlp.stanford.edu/projects/glove/

Table 4.6: Performance evaluation on SciCite [31] (intent) dataset. L = Layer, F = Filter, C = convolution size. Taken from [1].

Topography	Architecture	Class-based accuracy			micro-f1	macro-f1
		Result [%]	Method [%]	Background [%]		
CNN	L 3 F 100 C 3,4,5	79.92	76.53	79.24	78.50	78.56
CNN	L 3 F 100 C 2,4,6	81.85	77.69	81.14	80.12	**80.22**
CNN	L 3 F 100 C 3,3,3	64.09	71.74	85.46	78.05	73.76
CNN	L 3 F 100 C 3,5,7	76.45	74.05	85.46	**80.49**	78.65
CNN	L 3 F 100 C 3,7,9	68.34	70.58	87.26	79.20	75.39
LSTM	L 2 F 512	73.75	73.55	79.54	76.80	75.61
LSTM	L 4 F 512	75.29	69.59	82.95	77.54	75.94
LSTM	L 4 F 1024	68.73	70.91	84.25	77.75	74.63
RNN	L 2 F 512	25.10	56.86	62.19	55.30	48.05
BERT [35]	Base	84.56	75.37	89.47	84.20	83.13
ALBERT [67]	Base	83.78	77.03	87.06	83.34	82.62
ImpactCite [1]	Base	92.67	85.79	88.34	**88.13**	**88.93**
BiLSTM-Att [31]	*	*	*	*	*	82.60
Scaffolds [31]	*	*	*	*	*	84.00
BERT [13, 35]	Base	*	*	*	*	84.85
SciBert [13]	*	*	*	*	*	85.49

Table 4.7: Performance: Citation Sentiment Corpus (CSC). Taken from [1].

Topography	Modification	Class-based accuracy		
		Positive [%]	Negative [%]	Neutral [%]
CNN	*	28.2	21.3	94.8
CNN	Focal	36.9	16.9	94.3
CNN	SMOTE	39.4	20.2	84.2
CNN	Upsampling	36.1	6.7	92.8
LSTM	*	32.8	12.4	93.9
LSTM	Focal	42.7	19.1	82.8
LSTM	SMOTE	42.3	20.2	83.7
LSTM	Upsampling	26.1	11.2	**97.0**
RNN	*	24.5	21.3	72.7
BERT [35]	*	38.6	20.4	96.4
ALBERT [67]	*	44.3	28.8	95.8
ImpactCite [1]	*	**78.9**	**85.7**	75.4

4.4.2.1 *Experiment 1: Fixed Dataset Split on CSC Sentiment Dataset.*

For this experiment, we employed a fixed 70/30 data split for the CSC dataset excluding any additional dataset cleansing. We evaluated the performance of each previously used model. Additionally, we employed several sample strategies i.e. focal loss, SMOTE & upsampling, and analyzed their impact on the imbalanced data.

4.4.2.2 *Results and Discussion:*

The results of this experiment are shown in Table 4.7. We observed that all models mainly captured the concept of neutral citations. Additionally, we also observed that the methods like focal loss and SMOTE sampling increased the performance of the CNNs and LSTMs. Furthermore, upsampling does not help to improve the performance of the model. However, ImpactCite [1] effectively learned representations of each class. Especially, the negative class was captured in a much better way by ImpactCite. Although ImpactCite showed slightly worse performance on the neutral class, it performed significantly better for positive and negative classes. We conclude that ImpactCite can deal with the large class imbalance and show that the complex language models are superior to the baseline approaches enhanced with sampling and focus strategies for the CSC dataset.

4.4.2.3 *Experiment 2: Cross-Validation on CSC-Clean Sentiment Dataset.*

To compare our proposed ImpactCite with the results of Athar [8] we used a 10-fold-cross validation. However, due to the missing split information and the duplicates that exist in the original CSC dataset, we decided to experiment on the CSC-C dataset. Although the results are not directly comparable, the approach [8] is favored due

Table 4.8: Cross validation performance: Sentiment citation corpus (CSC-C)

Topography	**Class-based accuracy**			**micro-f1**	**macro-f1**
	Positive [%]	**Negative [%]**	**Neutral [%]**		
CNN	40.2	24.9	95.0	88.6	43.4
LSTM	34.8	19.0	92.1	84.6	46.1
RNN	20.7	17.9	86.0	77.9	41.5
BERT [35]	72.8	80.2	70.3	74.4	74.4
ALBERT [67]	71.1	72.5	67.6	70.4	70.4
ImpactCite [1]	64.6	86.6	82.0	77.7	**77.7**
SVM [8][5]	*	*	*	**89.9**	76.4

to the duplicates that appear in the training and test data. For the sake of completion, we included [8] as a reference. During the 10-fold cross-validation, we used nine splits as training and one split as a test dataset for each run and averaged the results at the end. A collection of experiments were performed employing a variety of models ranging from baseline CNN models to complex BERT language models. To successfully apply the baseline methods, we used the class weights as they have shown superior performance in previous experiments.

4.4.2.4 *Results and Discussion:*

The results of this experiment are shown in Table 4.8. Interestingly, the baseline models were not able to achieve comparable performance even though the class weights were employed. To resolve the class imbalance issue, we pre-processed the folds for the baseline approaches such that the number of positive and neutral training samples was decreased to the number of negative samples. Doing so resulted in the performances shown in the table. Additionally, we observed that the complex language models performed much better on the small dataset. They significantly outperformed the baseline methods and achieved good results across all three classes. In addition, ImpactCite [1] outperformed all other selected models and sets a new state-of-the-art for citation sentiment classification on the CSC-Clean. For the sake of completeness, we included the SVM used by Athar [8] evaluated on the CSC dataset.

4.4.3 *Out-Domain: Evaluating Impact of Additional Data*

In this section, we present our results using out-domain data to evaluate its impact on the model performance. We investigate multiple scenarios of cross dataset training and testing on datasets from different domains. Furthermore, we conducted experiments concerning the use of multiple datasets and an optimal schedule strategy to enlarge

[7] Trained and tested on CSC

the corpus size. We also discuss details of some experiments related to different data feeding methods.

4.4.3.1 *Experiment 1: Out-domain Testing.*

In this experiment, we employed a pre-trained XLNet for each dataset and fine-tune it on one dataset. Once the model is trained, we evaluated its performance across all datasets to find out which datasets are semantically closer to each other. The goal is to better understand the correlation of the dataset and to what extent it is possible to use the model trained on an out-domain dataset for the prediction of sentiment across other domains. In this experiment, we trained each model for 40 epochs with a batch size of 24. In addition, we also used an early stopping mechanism such that if the model converges before 40 epochs then it will stop further training to prevent over-fitting. It has to be mentioned, that in this experiment the datasets had different sizes, as shown in Table 3.1.

4.4.3.2 *Results and Discussion:*

In Table 4.9 we show the results when using a single training set and testing across all datasets. Overall the best performance was achieved using the same dataset for training and testing, the only exception is the Stanford dataset. Interestingly, the performance for the Stanford dataset is surprisingly good when the model is trained on the Cornell data. It has to be mentioned, that both datasets are from the same domain. This shows that training on more domain data without fine-tuning on a specific dataset can result in a pretty good model for that dataset which is taken from the same domain. Overall training on the Stanford dataset was not successful. In general training on a dataset of the same domain without fine-tuning the model resulted in a good performance on their domain however it is not the case when trained on out-domain data. One reason for this is the correlation between the data within the same domain. The results further show that the correlation across domains is in general lower but in the case of the Instruments dataset, the correlation is high enough to achieve superior performance using a dataset that is more balanced from the movie review domain. This suggests that a correlation between movie reviews and instrument reviews (product reviews) exists. Intuitively, this is the case because the understanding of positive and negative in the scientific domain is fundamentally different compared to reviewing data or tweets.

4.4.3.3 *Experiment 2: Sequential Training.*

In this experiment, we evaluated the impact of a sequential training scheme. The idea is that if a dataset is very small and therefore it is not possible to train only on that dataset, we enhance the data size by

Table 4.9: Results for testing on out-domain data using XLNets trained on a single dataset. Results are macro f1-scores in percent.

Test	**Movie**			**Product**	**Twitter**		**Scientific**
Train	IMDB	Cornell	Stanford	Instruments	Us Airline	Sentiment140	CSC-Clean
IMDB	**94.38**	81.58	83.66	70.20	64.53	62.72	54.16
Cornell	92.05	**89.69**	**94.39**	57.46	87.69	69.15	60.28
Stanford	91.71	89.49	92.85	63.68	86.46	68.89	63.76
Instruments	86.51	55.53	57.14	**82.73**	52.63	57.52	49.71
US Airline	56.80	71.45	79.47	43.80	**92.21**	68.39	43.45
Sentiment140	79.28	72.63	76.95	65.13	77.14	**80.57**	62.42
CSC-Clean	85.04	62.91	62.79	64.60	63.88	62.13	**76.67**

using additional datasets. There are two interesting aspects to using additional datasets. One it will increase the amount of data available for training and secondly, we also want to evaluate the impact of the dataset sequence in which data is fed to the network. Intuitively, the last dataset category in the training sequence should be favored with respect to the performance as the gradients are optimized on it. We performed this for a fixed sequence of datasets and categories and used several permutations of the sequence of categories to have comparable results. In addition, we performed these experiments twice, once for the upsampled datasets and once for the downsampled. The reason for this procedure is that it is important to make all datasets the same size such that they can contribute the same amount to the training. With the initial dataset sizes, this would not be the case and a few datasets would dominate the training due to their size. In the upsampled version we used $3,000$ samples whereas for the downsampling experiment we used the number of instances of the smallest dataset as a reference number. For some datasets, this means we had to select a subset of the training instances. This means we do not preserve the individual class distribution. The sequence of the datasets is shown in the corresponding results tables.

4.4.3.4 *Results and Discussion:*

In Table 4.10 we present the results for sequential training. The upper part of the table covers the training results using the upsampling whereas the lower part covers the downsampling results. Our results for the upsampling showed that putting the movie review data at the end achieved the best scores for three out of the seven datasets. The performances overall were superior to the scores of the downsampling. Using the movie data as the last dataset in the training resulted in a 78.18% macro f1-score for the scientific data which is 1.21% better compared to setting the scientific data at the end of the sequence. The downsampled part shows that the training with the product data, in the end, has shown the best performance for the three datasets. Interestingly, the performance on the scientific data was 8.21% better using the downsampled either the product or movie datasets in the end compared to using its own dataset as last in the downsampled scenario. Except for the testing on Instruments and the CSC-Clean dataset, the performances of the other datasets did not change dramatically based on the feeding sequence. Another interesting finding was that putting the movie reviews in the end for the downsampled experiments did not result in a bad performance for all other dataset categories and led to a maximum drop of 2.86% for the US Airline dataset compared to the best performance for that dataset. In general, it was not the case that the models shows a bias toward the dataset that was used last in the training epoch. It is to be noted that due to the computational effort we did not try every combination

Table 4.10: Macro f1-scores for sequential training. Sequence within the categories: [P]roduct (Instruments), [M]ovie (Cornell, IMDB Stanford Sent.), [S]cientific (CSC-Clean), [T]witter (Sentiment140, US Airline). 'Up' corresponds to the upsampled training data and 'Down' to the downsampled training data.

Test	**Movie**			**Product**	**Twitter**		**Scientific**
Train	IMDB	Cornell	Stanford	Instruments	Us Airline	Sentiment140	CSC-Clean
Up							
S T P M	**93.05**	**88.51**	90.87	80.22	89.69	75.45	**78.18**
M S T P	92.94	86.98	89.35	**80.25**	86.97	**77.16**	69.16
P M S T	91.62	87.81	90.05	74.32	89.84	76.25	70.39
T P M S	92.19	88.19	**91.26**	77.72	**90.04**	76.08	76.97
Down							
S T P M	**92.38**	**87.29**	**89.98**	**80.45**	85.93	**76.96**	**75.55**
M S T P	92.26	85.65	88.27	78.69	88.38	76.13	**75.55**
P M S T	90.55	85.94	89.27	65.93	**88.79**	75.33	66.73
T P M S	88.95	83.00	86.98	72.07	87.11	75.18	67.34

but selected a subset that puts every category once at each position. Furthermore, the general finding of this experiment series is that unexpectedly the network does not work better when trained last on the evaluating dataset. Although most of the achieved accuracies are comparable it is not easy to predict which sequence works best for which testing set. Generally, upsampling was superior for most of the datasets. However, it requires much more training time. In our case, the dataset size is $3,000$ compared to 797 samples for the downsampled version.

4.4.3.5 *Experiment 3: Shuffled Training.*

In addition to the sequential data feeding experiment, we performed similar experiments by shuffling the data. The major difference compared to the previous experiment was that there is no sequence preserved, neither within the categories nor between the categories. Therefore, the gradients can align to each of the data samples and are not biased towards the last category in the setup.

4.4.3.6 *Results and Discussion:*

In Table 4.11 we show the results of the shuffled upsampling and downsampling experiments. Surprisingly, the macro-f1 scores are close to each other. In these experiments, the downsampled data used about 800 instances of each dataset whereas the upsampled $3,000$. Even more interesting is that the shuffled model performed well across all datasets. The largest accuracy drop compared to the single dataset training models was about 3.44% for the Sentiment140 dataset. Comparing the performances of the downsampled model to the models trained exclusively on those datasets, the accuracy of the shuffled model is impressively good. The same holds for the upsampled model. In general, the shuffled model holds a better generalization as it can be applied to all the datasets even without fine-tuning and sticks to good performance.

4.4.4 *Multi-task Model: Fusing Scientific Sentiment and Intent*

4.4.4.1 *Experiment 1: Multi-Domain Usage*

We further experimented with the unified model for the sentiment and intent classification. This experiment combines both tasks into a single model. The motivation behind this experiment is to handle the increased amount of computation resource and inference time when using two separated models as proposed in ImpactCite. However, due to the small size of the CSC-Clean dataset, it is not possible to train it directly in conjunction with the intent task. Therefore, we utilized the previous findings and combined the citation sentiment data with

Table 4.11: Macro f1-scores for shuffled training. ’Up’ corresponds to the upsampled training data and ’Down’ to the downsampled training data.

Test / **Train**	**Movie**			**Product**	**Twitter**		**Scientific**
	IMDB	Cornell	Stanford	Instruments	Us Airline	Sentiment140	CSC-Clean
Up	93.65	88.04	91.81	88.90	89.99	77.13	74.45
Down	97.80	87.07	88.42	83.40	86.96	76.65	73.73

Table 4.12: Macro f1-scores sentiment and intent classification. Shows that the single task model is superior for the individual tasks.

Setup / **Task**	**Mutli-Task**		**Single-Task (ImpactCite)**	
	All sent. datasets	CSC-Clean + Stanford	CSC-Clean	SciCite
Sentiment	64.00	56.00	**80.41**	∗
Intent	78.00	78.00	∗	**88.93**

the sentiment datasets from other domains to enlarge the training set. Therefore, the sentiment task covers the sentiment classification for all used datasets that included a neutral class.

4.4.4.2 *Results and Discussion:*

Results in Table 4.12 show that the unified multi-task model has advantages however it is achieved with certain limitations. Firstly, the advantage of the multi-task model is that only a single model is used and two different heads are trained. This makes inference twice as fast as only one forward pass is needed and reduces the required hardware. However, the only impediment is that the model is trained on the conjunction of sentiment data and therefore the bias of the out-domain context can hinder the intent performance. It is to be noted that the model is robust against out-domain data for the sentiment task.

4.5 DISCUSSION

In the first section [1] we have shown that our approach is capable to perform well on both the sentiment and intent classification. The results highlighted the problems with the scientific sentiment domain and the lack of data. Additionally, the unbalanced datasets resulted in difficulties to converge for all evaluated methods except ImpactCite [1]. Neither ALBERT [67] nor BERT [35] was able to converge up to a state that provides a sufficient performance across all tested classes. While an intent classification using those models works well. However, this is not the case for sentiment classification as some classes were not captured by the models. Especially, the negative class was identified as one of the major shortcomings. However, we were able to overcome this data shortcoming up to a certain extent using ImpactCite [83]. We achieved a new state-of-the-art performance for both tasks emphasizing the gains using XLNet [129] when the existing data is limited and unbalanced. In addition, these findings served as a baseline for qualitative citation analysis which is most times not considered due to the lack of available datasets.

In this chapter, we mainly focused on the utilization of out-domain data to enhance the sentiment classification in the scientific domain which suffers from the lack of existing annotated datasets. Our experiments have shown that without a specific fine-tuning the correlation between in-domain datasets is stronger compared to out-domain datasets, and it is possible to achieve surprisingly good results training a classifier on a dataset of the same domain even without fine-tuning. Interestingly, in some cases, the larger quality datasets have shown better performance on some test sets than using the original training set. Going one step ahead, we evaluated different scheduling techniques to better understand the impact of data fusion.

First, we tried different sequential concatenations resulting in better-generalized models that we can perform well across all datasets. Although the sequence has been shown to bias the performance slightly towards the last category the results showed that the movie data as the last set in the sequence performed best. In addition, the difference between the upsampled and downsampled training dataset versions highlighted that if the number of datasets concatenated is sufficient then this approach works for very small datasets below 800 samples. Next, we mixed all sentiment training data to avoid preserving sequence to favor any of the domains which resulted in a superior model with respect to the generalization. Shuffling all the data removed the convergence towards a single domain. Although it would be possible to fine-tune the model on a single dataset. We demonstrate that our solution is more robust as it is confronted with out-domain data during the training and further utilizes this data to establish a more general understanding of the underlying language concepts that are not bound to one domain.

Ultimately, the combination of tasks within a single model can be very complex. During our experiments, we faced several challenges while combining the sentiment and intent tasks. It was not possible to train a model that is capable to converge using only the scientific sentiment and intent data. This is the case as the sentiment data is very small and when combined with the intent task, the network is not able to learn the concept of sentiment, especially negative sentiment, due to a large amount of unrelated data. Although we have shown in our previous work [1] that the use of two separate models is possible this might not be desired as the hardware required to run two models parallel is expensive. Furthermore, a sequential inference suffers from time delay. As a feasibility study, we combined the sentiment data with the out-domain sentiment data and trained the multi-task model. Ultimately, the proposed model is capturing multiple tasks and domains.

5 SEMANTIC INDEX

The assessment of the scientific research conducted by a researcher and its impact is a crucial part of every researcher's career. The scientific ranking of a researcher can be decisive for the hiring decisions of universities and institutes as well as for the allocation of research funding and represents the prestige among the scientific community. Researchers generally publish their findings in the form of a scientific publication therefore, they would be referred to as authors from here onward in this chapter.

Bollen et al. [18] describe science to be a gift economy. Within it, the value of an author can be defined as the degree of his contribution to knowledge as well as the degree of how much he impacts the ideas of other scientists. As Hirsch [55] pointed out, it is needed to have the possibility to quantify this kind of value, among others, for recruitment decisions of universities and the award of grants, especially in a world of limited resources. The increasing costs of research and the shortage of available economic resources lead to a high and increasing interest in scientific author assessment [32]. Additionally, the usefulness of evaluating scientific author impact and author ranking when doing research, in general, should not be underestimated. It offers the possibility for every researcher to easily spot authors heavily contributing to a research field and to discover their publications which might be worth reading when executing research in a specific field. For achieving such an author impact assessment, different indicators are commonly used by many author assessment approaches. On the one hand, there are production indicators which are, for example, the total number of published papers and on the other hand, there are impact indicators which are usually based on the citations received by an author [5]. Hirsch consequently states that the large amount of useful information, which is given by the publication record of an individual, can be evaluated with different criteria by different researchers [55]. This leads to the emergence of different author assessment approaches. Each of these approaches can be considered as an attempt to highlight a specific aspect of an author's publication record that might be of interest when evaluating the author's importance and contribution to science [24]. There are huge debates on which of them are the best for assessing the importance and contribution of a scientific author. However, it is widely accepted as a good approach to simply use multiple quantitative measures to support an expert judgment for improving objectivity and fairness in the evaluation process.

The contributions of this chapter include an overview of some existing authors' research assessment indexes to evaluate and measure the impact of scientific authors in their research field, followed by discussing the limitation of each index. However, the main contribution of this chapter is to propose a new author's research assessment index which takes more semantic and qualitative aspects into account and could meaningfully represent the impact of scientific research work.

Author indexes are generally divided into two categories. Classic and Weighted author indexes, where each category has a range of proposed indexes belonging to it. We will discuss the details of each category and their respective approaches in the following sections.

5.1 CLASSIC AUTHOR INDEXES

Classic author indexes take into account raw indicator values to quantify the impact of an author's research. The most common way to estimate a classic author index is to select an indicator i.e., citation count. Based on the selected indicator, scientific author indexes can be formulated, which offers a standardized way for measuring author impact and suggests a high or low rate of impact of an author's research. These classic scientific author indexes differ from the weighted indexes in the way that they do not offer individual weighting for each citation which is used for estimation of the index value. Consequently, classic author indexes are unable to handle the major problems of self-citation and co-authorship. Following are some popular examples of the classic author indexes:

5.1.1 *h-index*

The h-index is one of the most commonly used scientific author indexes. It was introduced by Jorge E. Hirsch [55] as a measurement to characterize the scientific output of a researcher in a simple and useful way. The h-index of a researcher represents the maximum value of h, where h stands for the number of publications by a researcher that is cited at least h times. If 5 papers of a researcher are cited at least 5 times each, then the h-index of that researcher is 5. Moreover, the publications considered for estimating the h-index are known as Hirsch core or h-core.

Based on this definition, the h-index has some advantages compared to other single-number indicators such as, simply using the total number of published papers or citations. By combining an indicator for the activity of publishing papers and an indicator i.e. citations for the scientific impact, it can be assumed the h-index is a relatively robust index. Publishing more papers does not influence the h-index. Similarly, a single publication with a significantly high or low number of citations does not affect the h-index of that researcher. In general, it can be inferred that the h-index is insensitive to outliers in both directions [32, 59].

On the other hand, these arguments also uncover an important limitation of the h-index. For Instance, new researchers might have a hard time catching up with the h-index value of already longer active scientists. This is because the h-index depends only on the number of published papers and the number of citations received by those papers, which are initially equal to zero for new researchers. Even if the first paper of a researcher has a substantial impact in a specific research field, the h-index of the researcher stays at $h = 1$, irrespective of the fact that how often his work is cited and how large the contribution to the research field is made by their first paper. Consequently, comparability between authors on the h-index can only be given in the long-term observation of their work. Additionally, the h-index has no mechanism to avoid the unsubstantial use of self-citations to increase the personal h-index. Moreover, the h-index value of an author can mathematically never decrease. The h-index of an author tends to increase without any additional contribution from the researcher because of additional citations over time. This results in researchers who started publishing earlier, always having an advantage over the researchers who started publishing more recently, even if the first-mentioned no longer contribute substantially to the research field.

Following the proposal of the h-index by Hirsch, many researchers tried to overcome the limitations of the h-index. In the following years, many more indexes similar to the h-index were introduced by improving and modifying the original h-index [59].

5.1.2 *g-index*

In 2006 Leo Egghe proposed the g-index [37] as an improved version of the h-index, just one year after Hirsch proposed the h-index. According to Egghe's definition, the g-index represents the number of papers g, where all g papers received in total g^2 or more citations. This leads to the fact that the g-index is always higher than the h-index. It also adds more variance in the values of the index for scientists within a research field and makes the g-index more precise to compare the authors with each other regarding their contribution to the specific field.

Costas et al. [32] performed a study to discuss the advantages and limitations of the g-index in particular compared to the original h-index. They found out that the main advantage of the g-index is that it pays attention to the weight of the citations received by the top publications of a researcher. They highlight that the g-index value is not limited by the total number of papers published by a researcher and hence favors the researchers who publish a low volume of publications with an exceptionally large impact in their research field.

Researchers following this publishing strategy are discriminated by the original h-index because their index value is limited by the small number of published papers and ignores the exceptionally high number of citations on these few documents.

The study concluded that the g-index overcomes some problems of the h-index while having its critical limitations. They argued that the higher sensitivity leads to the g-index specific limitation of the disproportionate impact of single highly cited documents, which can be seen as outliers regarding the number of citations a specific author receives in general. The g-index also shares some limitations with the original h-index, which are, for example, the problem of self-citation and the missing ability to evaluate the contribution of scientists across different research fields. Additionally, they pointed out that the two indexes should be seen as complementary and not substitute each other.

5.1.3 *hg-index*

Inspired by the h-index and the g-index, Alonso et al. [5] introduce the hg-index in 2010. To prove the need for their hg-index, they explicitly reference Ronald Rousseau [105] who argued on the role of the h-index and the g-index by highlighting the limitations of the h-index and dismissing a more sensitive g-index as a replacement for the h-index. They emphasized using both indexes in conjunction. Alonso et al. realized this idea and proposed the hg-index, which represents the geometric mean of an author's h-index and g-index. It can be mathematically expressed as

$$hg = \sqrt{h \times g} \tag{5.1}$$

Where h denotes the author's h-index and g denotes the author's g-index. Additionally, they state that $h <= hg <= g$ as well as $hg = h <= g = hg$ and that the hg-index corresponds to a value nearer to the value of the h-index than to the value of the g-index. In the case of a very low h- index this can be seen as penalization of the g-index.

Using the combination of the h-index and the g-index, it is possible to acknowledge highly cited papers, as the g-index would do, while still significantly reducing the impact of single very highly cited papers, as the h-index would do. Additionally, it is very simple to calculate the hg-index when the h-index and the g-index of an author are already known. The simple calculation as a geometric mean also leads to the advantage of having the same scale for all three indexes, which gives the possibility to compare the values across these indexes.

Besides the advantages the hg-index has to offer, it should be mentioned that the proposal of the hg-index also received criticism. Franceschini et al. [43] pointed out that in their opinion, the hg-index has the following limitations:

1. hg derives from a composition of two indicators, h, and g, defined on distinct ordinal scales
2. the equivalence classes of hg are questionable and the substitution rate between h and g may arbitrarily change depending on the specific h and g values
3. the apparent increase in granularity with respect to h and g is illusory and misleading

5.1.4 *A-index*

Inspired by the idea of the g-index, Jin et al. [59] proposed A-index. It is defined as the average number of citations received by h-core publications. A-index overcomes the h-index limitation of ignoring the exact number of citations received by publications within the h-core. Mathematically, it can be expressed as

$$A = \frac{1}{h} \sum_{j=1}^{h} \mathtt{cit}_j \tag{5.2}$$

Where h denotes the author's h-index and $\mathtt{cit}_j$ represents the number of citations the elements within the h-core receive in decreasing order. With this definition in mind, the A-index suffers from a major problem: Independent of the number of citations on the documents within the h-core, authors with a higher h-index are discriminated by their A-index value compared to authors with a lower h-index because the A-index definition contains a division by h.

This problematic behavior of the A-index can be vividly described with an example case presented by Jin et al. [59] where there are two researchers A and B. Researcher A has 20 publications where one publication is cited 10 times and all other publications are cited only once. On the other hand, Researcher B has 30 publications where one publication is cited 10 times and all other publications are cited twice. It results in the h-index of researcher A and B as 1 and 2 respectively. However, their A-index values are 10 and 6 respectively due to the division by h while estimating A-index. Therefore, A-index seems to penalize the authors with a high h-index.

5.1.5 *R-index*

In order to address the problems in A-index, Jin et al. [59] proposed R-index which leaves out the division by h and uses the square root of the number of citations received by all documents within the h-core. Mathematically, the R-index is defined as:

$$R = \sqrt{\sum_{j=1}^{h} \mathrm{cit}_j} \tag{5.3}$$

Where h denotes the author's h-index and cit_j represents the number of citations the elements within the h-core receive in decreasing order. It can be noted that $h <= R$ because each cit_j is at least equal to h. The formula also leads to $R = h$ for the case that cit_j is exactly equal to h [59]. However, R-index has its inherent limitation as it may be overly sensitive to one publication of an author receiving a very high number of citations.

5.1.6 *AR-index*

One of the limitations of the h-index mentioned in section 5.1.1 is that the h-index value of a researcher can never decrease mathematically. This leads to scientists being able to rest on their laurels. By proposing AR-index, Jin et al. [59] tried to mitigate this problem. The AR-index is represented as

$$AR = \sqrt{\sum_{j=1}^{h} \frac{\mathrm{cit}_j}{a_j}} \tag{5.4}$$

Where h denotes the author's h-index, cit_j represents the number of citations document j received and a_j denotes the age of document j. The definition shows that while summing up the number of citations received by the publications within the h-core, the AR-index adds a weighting based on the age of an article. This leads to the possibility of the AR-index decreasing over time if the publications within the h-core do not have a significant citation increase to compensate for the decreasing weighting due to their publication age [59].

While the AR-index solves a limitation of the h-index, it may not be appropriate to use the AR-index to evaluate the lifetime contributions of a person to a research field. It is so due to the above-mentioned devaluation of the citations received by the publications over time. Therefore, jin et al. [59] recommend always using AR-index together with the h-index to get complete context.

5.2 WEIGHTED AUTHOR INDEXES

In this section, we discuss the other type of index called the weighted author index, which focuses on the individual weighting of each citation instead of assigning weights to a set of selected citations or the overall citation count. These approaches are further divided into two categories i.e. citation weighted and network weighted. Both of these categories are discussed as follows:

5.2.1 *Citation-based Weighted Author Indexes*

In this category, the approaches focus on assigning weights to each citation based on either of the two aspects i.e. self-citation and co-authorship. Following are the solutions proposed for each of these aspects:

5.2.1.1 *Handling Self Citations*

In 2009, Schreiber et al. [111] studied the influence of self-citation on the h-index. In contrast to Hirsch's proposition, they proved that self-citations have a large impact on the h-index. They measured the decrease of the h-index to be $21,3\%$ by average when excluding self-citations from the index calculation for the dataset selected for the study. They argued that the significance of a publication is usually not reflected by self-citations. It led to the necessity to treat self-citations differently from regular citations. Therefore, they introduced the sharpened index h_s-index, which excludes self-citations from the index calculation and sets the weighting of self-citations to zero. It inspired the researchers in the domain who adopted this weight assignment method for their solutions i.e. ch-index [65] and b-index [23].

Schubert et al. [14] argued that it is not appropriate to simply exclude the self-citations and therefore proposed an approach called Fractional Self-citation Counting, which deals with handling citations for publications with multiple authors. It takes two factors into account i.e. the number of co-authors in citing publication and the number of co-authors in cited publication. They proposed to use the Jaccard Index to identify the overlap of authors incited and citing papers and based on the overlap assign weight to the citation. The Jaccard Index is represented as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{5.5}$$

Where A represents the set of authors of the citing paper and B represents the set of authors of the cited paper. The weight of the citation is decreased based on the amount of overlap of authors found in both cited and citing papers. It was argued to be more justified than simply excluding the complete citation for all authors.

5.2.1.2 *Handling Multiple Co-authors*

Having multiple co-authors also poses the same challenge as the self-citations while estimating author indexes. Sekercioglu et al. [113] studied the trends of co-authorship in the scientific community and pointed out the average number of co-authors is on the rise. There also exist papers with more than 500 co-authors. They argued that giving credit to all authors of a publication is not fair and emphasized devising an approach to have a standardized process of quantifying co-author contributions. They proposed that kth co-author in the author's list should be assigned a contribution of $\frac{1}{k}$ as compared to the contribution of the first author so that the contribution of all co-authors can sum up to one.

Zhang [130] criticized the solution presented in [113] by highlighting that it does not take into account the importance of the corresponding author, which is usually the last author in the author's list. To tackle this problem, Zhang [130] proposed the w-index which employs the idea of weight coefficients. These are the multipliers that regulate the assignment of the credit to an author of a paper in an author's list of a publication. Both corresponding authors would have a weight coefficient value of 1, while for the remaining authors the weight decreases linearly with the increasing rank. Galam [45] proposed a similar idea to assign more weight to the first and corresponding author. They employed Tailor Based Allocations (TBA) for weight assignment to all authors in a publication. Furthermore, they claimed that this approach represents an incentive for stopping the inflation of publications, which would be the result of giving every co-author full credit for a collaborative published paper.

According to Hagen [50], there are three basic ethical criteria for equitable sharing of authorship credit which are defined as:

1. One publication credit is shared among all co-authors
2. The first author gets the most credit, and in general, the ith author receives more credit than the $(i+1)$th author, and
3. The greater the number of authors, the less credit per author

These three criteria are all fulfilled with the so-called harmonic counting [51] they introduced as another, more complex, approach for dealing with the weighting of citations for cooperative published papers. When using the harmonic counting method, the authorship credit for one specific author of a cooperative published paper is defined as

$$\text{Harmonic } i\text{th author credit} = \frac{\frac{1}{i}}{[1+\frac{1}{2}+\ldots+\frac{1}{N}]} \tag{5.6}$$

Where i denotes the position of the author in the co-author list and N denotes the total number of co-authors of the paper.

5.2.2 *Network-based Weighted Author Indexes*

This section discusses the approach which generates networks using citations and collaboration data to identify the importance of a researcher in a scientific community. A few famous approaches are discussed as follows:

5.2.2.1 *Weighted PageRank*

The PageRank algorithm [22] was originally developed to evaluate the relevance of a web page. Yan et al. [128] proposed a modified implementation of the original PageRank [22] algorithm to assess the importance of authors and publications. They proposed to apply a weighted PageRank algorithm to a citation and co-authorship network for providing a measurement that gives a full perspective on the impact of an author. Yan et al. [128] modified and transferred a weighted version of this original PageRank algorithm into the context of measuring author impact. Mathematically, they defined their weighted PageRank index for assessing the scholarly impact of an author as

$$PR_W(p) = (1-d)\frac{CC(p)}{\sum_{j=1}^{N} CC(p_j)} + d\sum_{i=1}^{k}\frac{PR_W(p_i)}{C(p_i)} \tag{5.7}$$

Where $PR_W(p)$ is the weighted PageRank of author p and the weighted PageRank of another author p_i in the network who is citing author p is expressed as $PR_W(pi)$. Analogous to the original PageRank, N is the number of existing publications in the network, d is the damping factor, and $(1-d)$ is the coefficient to retain the sum of the PageRank as one. $CC(p)$ represents the number of citations received by author p and $\sum_{j=1}^{N} CC(p_j)$ represents the number of citations received by all authors in the network while $C(p_i)$ now stands for the number of outgoing citations from author p_i.

5.2.2.2 *Author-level Eigenfactor*

West et al. [124] present a way to adapt the Eigenfactor score, which was originally introduced for ranking journals, to use the author-level citation data as a base for ranking the scholarly output of authors, institutions, and countries. To estimate the Eigenfactor score of authors within a citation network, an iterated voting procedure was adopted. In the beginning, each author’s single vote is proportionally divided across all the authors cited by the author in question. Then the same procedure is repeated for the authors cited in the previous iteration and so on until a steady-state is reached. Eventually, the Eigenfactor score of an author is the sum of received votes by that author at the steady-state. The authors claim that the Author-level Eigenfactor helps the researchers for finding important papers, that may have been overlooked by other ranking methods.

5.2.2.3 *TimeRank*

Franceschet proposed an approach called TimeRank [42] which introduces citation timing as another aspect for the evaluation of scientific author rankings. The basic idea of TimeRank is to allocate the rating by considering the relative position of two authors at the time of the citation among them. In TimeRank, initially, all authors have the same rating i.e., 0. With time, more citations are processed. At any time $t > 0$, the ratings of scholars cited at time t are simultaneously updated in terms of their previous ratings at time $t - 1$ and the previous ratings at time $t - 1$ of the citing scholars.

The amount a citation of a specific author i improves the rating of the cited author j is noted as citation reward $p_i j$. It is defined that $0 < p_i j < 1$, which means that the citation reward is always positive because every citation, no matter from whom, should contribute to the ranking of the cited author. Further, it is defined, that the reward $p_i j$ is close to 1 if the citing author i is significantly higher rated than the cited author j and the reward $p_i j$ is close to 0 if the citing author i is significantly lower-rated than the cited author j. In the case that both authors are rated similarly, the citation reward is close to 0.5. TimeRank is more relevant in areas of application where the quantity, quality, and timing of the publications of a researcher are relevant.

5.3 SEMANTIC INDEX

In this section, we will discuss the formulation of proposed Semantic Index scores to assess the influence of the researchers in a scientific community. The purpose of the Semantic Index is to assign a representative score to individual authors which depicts the extent of their contribution and its acceptance in the scientific community.

Generally, author indexes take into account the quantitative aspect of citations i.e. Number of citations received by publications. We propose a new Semantic Index that considers the nature of individual citations in addition to citation count, therefore, enabling us to integrate the qualitative aspect of citations in the Semantic Index. In this work, we consider two qualitative aspects of citations namely citation sentiment and self-citation. The motivation behind this selection is fairly intuitive as we propose that not all the citations are equal, the first factor which sets apart one citation from the other is whether a paper is cited positively or negatively i.e. appreciating and using the proposed approach or highlighting shortcomings of a research work respectively. In Semantic Index, we only consider the citations which have a positive sentiment as those citations represent the appreciation and support of the scientific community for research work. In this work, we estimate the citation sentiment by using the approach mentioned in Chapter 4. The second factor which affects the quality of citation is whether an author is citing their own papers which is synonymous with a famous English idiom 'Self-praise is no recommendation'. Therefore, any citation which is an instance of self-citation is not considered during the estimation of the Semantic index for an author. The resultant number of citations will be referred to as $N_{positive}$ in this section.

In addition to the above-mentioned qualitative aspects, we also consider the multi-faceted community interactions of an author to effectively evaluate their position in a scientific community. These multi-faceted interactions are represented by different centrality measures. In graph theory, a centrality measure is used to rank nodes based on their position in the graph. To estimate these centrality measures, we use two types of graph networks one is the author citation network and the other is the author collaboration network.

In this work, we construct an author citation network by representing all citations extracted from the publications in the form of a directed graph where each node represents an author of a publication, and the relation between two nodes highlights citations pointing in the direction of cited author. On the other hand, the author's collaboration network depicts the collaboration among the authors in a scientific community. It is constructed using the information extracted from the header of a publication, where each author is represented by a node and a non-directed relation between two nodes represents collaboration in a publication. Once both author citation and collaboration networks are ready, we can now estimate the value of different centrality indicators for the semantic index of an author. Each of the centrality measures verily represents a role of an author in a scientific community. The description of the selected centrality measures is as follows:

1. **Degree Centrality:** Degree centrality of a node refers to the number of nodes connected to that node. In this specific context, we use the author collaboration network for computing degree centrality. The values of degree centrality are normalized by dividing with the maximum degree existing in the graph. It represents the extent to which an author is connected in a community network.

2. **Eigenvector Centrality:** Eigenvector centrality of a node in a citations network depends on the centrality of its neighboring nodes. For this purpose, we prepare an adjacency matrix A of all vertices V in the citations graph. The eigenvector centrality of a node i is the ith value in the normalized eigenvector. It can be represented as in equation 5.8, where A is the adjacency matrix, v is the eigenvector and λ is the constant representing the highest eigenvalue of A. The influence of each vertex in the network is iteratively estimated using the influence values of

its neighboring nodes resulting in an eigenvector with updated values. This process is repeated until the ratio between the values in the eigenvector converges. The resultant vector is then normalized such that the most important node will have the eigenvector centrality score of 1. Eigenvector centrality quantifies the transitive influence of a node on its neighboring nodes.

$$Av = \lambda v \tag{5.8}$$

3. **Betweeness Centrality:** Betweenness centrality refers to the influence of a node on the flow of information between distinct parts of a graph. For each node, it is estimated using the number of shortest paths between all nodes in the graph passing through that node. Equation 5.9 shows the formal representation of betweenness centrality where V refers to the nodes present in a graph, $\sigma(s,t)$ is the shortest path between the nodes s & t and $\sigma(s,t|v)$ refers to the shortest path between the nodes s & t which passes through node v. It measures the influence and control of a node on the flow of information.

$$B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{5.9}$$

4. **Closeness Centrality:** Closeness centrality of a node u is the measure of its average closeness to all other nodes in the graph. It is estimated by summing the shortest paths from the node u to all other nodes in the graph and later taking the reciprocal of this sum. The value of the closeness centrality can be normalized by multiplying it with $n-1$ where n is the total number of nodes in the graph. Normalized closeness centrality can be represented as the equation 5.10, where $d(v,u)$ is the shortest distance between nodes v and u. Closeness centrality estimates the extent to which a node can efficiently spread information.

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{5.10}$$

5. **Indegree Centrality:** Indegree centrality refers to the number of incoming connections to a node. This value of indegree centrality is computed using the citations network. The values of indegree centrality are normalized by dividing with the maximum indegree found in the graph. It represents the number of incoming connections i.e., citations from other nodes.

Each of these centrality measures refers to a specific role in the scientific community which would be discussed in the next chapter. After computing all the centrality measures for every node in the network, we prepare the weighted centralities by taking a product of non-self-cited positive citations $N_{positive}$ with the sum of all centrality scores to finally compute the Semantic index value for each node in the network. The proposed Semantic Index can be formulated as follows:

$$Index_n = \log((c_{deg} + c_{eig} + c_{bet} + c_{clo} + c_{ind}) \times N_{positive}) \quad (5.11)$$

Where $N_{positive}$ represents the total number of positive citations received by an author. It is to be noted that $N_{positive}$ does not include any self-citation. c_{deg}, c_{eig}, c_{bet}, c_{clo}, and c_{ind} represent the Degree, Eigenvector, Betweeness, closeness, and indegree centralities respectively.

5.4 EVALUATION

We evaluated the Semantic index by inspecting its compliance with Declaration on Research Assessment (DORA) guidelines which were developed in 2012 as a result of the Annual Meeting of the American Society for Cell Biology in San Francisco. These guidelines were further developed and refined over a period of time and are currently been actively maintained. This initiative provides instructions and best practices to all researchers, organizations, funding agencies, and scientific communities to assess the quality of scholarly research. So far, there are 21,729 participants and organizations from 158 countries who have already signed the DORA declaration. For our comparative evaluation, we selected 5 most widely adopted author indexes i.e. h-index, g-index, i10-index, Eigenfactor, and Impact factor.

5.4.1 *DORA Guidelines*

In this section, we will discuss the core aspects of the DORA guidelines and their compliance in the case of the Semantic index and some most common author indexes.

5.4.1.1 *Suitability for Quality Evaluation*

This aspect represents the overall purpose of an author index, which is to evaluate the quality of the research work performed by a researcher. However, in Section 5.1 and 5.2 we discussed different author indexes along with their advantages and limitations. It is evident from the discussed limitations that the h-index, g-index, i10-index, and Eigenfactor are not suitable for evaluating the quality of research work as they heavily rely on to even consider a publication for assessment. For instance, any publication with citations less than the h-index of an author will be discarded. The impact factor was initially introduced to help librarians to facilitate in deciding which potential journal volume they should buy for their libraries and now the scientific community seems to measure the quality of a journal using the impact factor. Such affairs make the existing indexes unsuitable for accessing the quality of research work. However, our proposed Semantic index considers the qualitative aspect of a publication i.e. citation sentiment to justify its suitability to serve as a tool for assessing the quality of research work.

5.4.1.2 *Circumstances*

The second aspect that DORA deems important for evaluating the impact of research is the consideration of individual circumstances. For instance, a researcher who joined recently would have less time to get citations as compared to the long publishing old researchers. All h-index, g-index, i10-index, Eigenfactor, and Impact factor do not take into account such individual circumstances of an author and usually take a very long time to gradually increase the score of these indexes. On the other hand, the Semantic index considers all publications irrespective of their number of citations and hence provides a consistent increase in score upon receiving new citations.

5.4.1.3 *Content Oriented*

Another aspect of impact assessment of research work is to consider the content of the publications while performing the assessment. As already mentioned, the h-index, g-index, i10-index, Eigenfactor, and Impact factor only consider raw citation count to estimate the impact of research work and all these indexes do not consider the content of the publications. Citation count is a superficial feature, as it does not covey any information about the sentiment of a citation i.e. if a publication is cited positively, negatively, or neutrally. Contrary to this, the Semantic index takes into account the content of the publications to identify the sentiment behind a citation so that it can be given credit accordingly.

5.4.1.4 *Protection against Manipulation*

Robustness is another important aspect of a quality index. Due to the high dependency of the h-index, g-index, i10-index, Eigenfactor, and Impact factor on raw citation count, it enables the malicious actors to manipulate their index scores with relative ease. The number of citations keeps on increasing without considering their source resulting in manipulation of index values. The selected indexes have unfortunately no mechanism to tackle actions like self-citations or a closed

Table 5.1: Aspects Covering DORA Guidelines for Evaluating Research Impact.

Aspects	**h-index**	**g-index**	**i10-index**	**Eigenfactor**	**Impact Factor**	**Semantic Index**
Quality	−−	−−	−−	−−	−−	++
Circumstances	−−	−	−−	−	−	++
Content Oriented	−−	−−	−−	−−	−−	++
Manipulation	−−	−−	−−	−	−−	+
Reliability	−	−	−	−−	−−	++
Transparency	+	+	+	+	−	++

loop of citations. On the other hand, Semantic Index is relatively robust as it discards all the self-citations while estimating its value for a given researcher.

5.4.1.5 *Transparency & Reliability*

Transparency & Reliability of indexes are the key aspects of analyzing the quality of research work. Existing indexes are somewhat transparent as we know that based on what data it was estimated. However, they have limited reliability as the publications with a low number of citations are completely overlooked which might be crucial in some domains i.e., Medicine. In contrast, for estimating the Semantic index, data is collected directly from the publications, therefore it is much more transparent and reliable. Additionally, it is generalizable as it considers all publications with any number of citations thus making it suitable for any domain.

5.4.2 *Comparison with Existing Indexes*

In this section, we will compare the limitations of existing indexes with the Semantic index. Similar to the previous section, we have selected the most widely used indexes for comparative analysis i.e.h-index, g-index, i10-index, Eigenfactor, and Impact factor. Following are the key limitations of all the existing indexes:

5.4.2.1 *Self Citations*

Self-citation is the most common challenge when assessing the impact of a research profile. Indexes like h-index, g-index, i10-index, and Impact factor do not handle self-citations and continue to consider them during the estimation of their index values. However, unlike other indexes, Eigenfactor and the Semantic Index discard the self-citations as they are not considered to be a part of the actual impact of a research publication.

5.4.2.2 *Lack of Fairness*

Indexes like the h-index and g-index favor the old researchers who have been publishing for a while to sustain their index score as their publications received many citations over years or decades. For the new researchers, they have to wait for years or decades to accumulate a high number of citations and eventually reach the same level as old researchers. On the other hand, the i10-index, Eigenfactor, and Semantic index consider the citations from all papers, and their values start increasing with the growing number of citations. Therefore, it does not require a lot of time to build up, hence supporting new and existing researchers relatively fairly & equally.

5.4.2.3 *Quantity vs Quality*

In the context of publishing, the scientific community has two popular schools of thought. One focuses more on the quantity of publications and the other one emphasizes more on the quality of the publications. This results in the cases where researchers have either High volume and low quality or low volume and high quality respectively. The selected indexes are ineptly not able to handle such cases as either of these cases effect the index score of the h-index, g-index, i10-index, Eigenfactor, and Impact factor. Contrary to popular indexes, the Semantic index can well handle the delicate balance between Quantity and Quality. Since the Semantic index favors all publications equally and therefore can handle both of these cases.

5.4.2.4 *Coverage*

Some indexes i.e. h-index and g-index employ components like h-core and g-core which results in partial coverage of publications including citations. These limited components restrict the scope of insights provided by the index scores and hence provide an incomplete picture of a researcher in the community. On the other hand, Eigenfactor, Impact factor, and Semantic Index are independent of typical h-core or g-core components to estimate the index score for a given researcher. Therefore, the limitations associated with h-core or g-core are irrelevant for these indexes.

5.4.2.5 *Meaningfulness*

Author indexes like h-index, g-index, i10-index, and Eigenfactor are meaningful to some extent as they attempt to highlight the importance of notable authors in the scientific community. However, they severely lack evaluation of the quality of the work due to the several limitations mentioned above. This results in partial meaningfulness of these indexes. On the contrary, the Semantic index not only analyzes all available citations but also the quality of each citation by analyzing the sentiment of each citation. The semantic index has the same granularity as most popular indices like the h-index. However, the Semantic index is more comprehensive and is, therefore, more meaningful.

Part IV

DATA VISUALIZATION

6

ACADEMIC COMMUNITY EXPLORER

In this chapter, we propose a comprehensive system called Academic Community Explorer ACE 2.0 for analyzing the scientific communities. It is a modular system where each module is responsible for carrying out a specific task. For each module, we employed state-of-the-art models for the tasks like bibliographic reference detection, citation sentiment analysis, and keyword detection & topic modeling. Data from each of these modules is collected and consolidated to perform the digital profiling of all the involved entities i.e., researchers and communities. The consolidated data is then analyzed to identify trends in interactions and the computation of statistics. Once the analysis is finished, all the findings are then visualized using a visualization engine. With the help of different examples discussed in this chapter, we can get instant community insights. These insights help us in identifying the practices employed by the authors, their citation patterns, and the interests of the scientific community. These findings later assist us in policing and paving the path for the future direction of a scientific community. Fig. 6.1 shows an overview of ACE 2.0 System. The pipeline consists of three stages. Each stage specializes in data extraction, analysis, or visualization. A stage may consist of multiple modules where each module performs a specific task. Details of each stage are discussed as follows:

6.1 STAGE1: DATA EXTRACTION

This stage is responsible for extracting data from given scientific publications. There are three modules in this stage and are responsible for identifying and extracting bibliographic references from scientific publications, identifying the keywords and topics associated with each publication, and analyzing the sentiment of a scientific publication. We employ our work from previous chapters as individual modules to perform these tasks. Each of the selected modules are already discussed in detail in Chapter 2, 3 and 4 respectively.

6.2 STAGE2: PROCESSING & ANALYSIS

As the name suggests, this stage is responsible for processing the collected data and analyzing it to identify different trends which provide useful insights into a scientific community. There are three modules in this stage. Details of each module are discussed as follows:

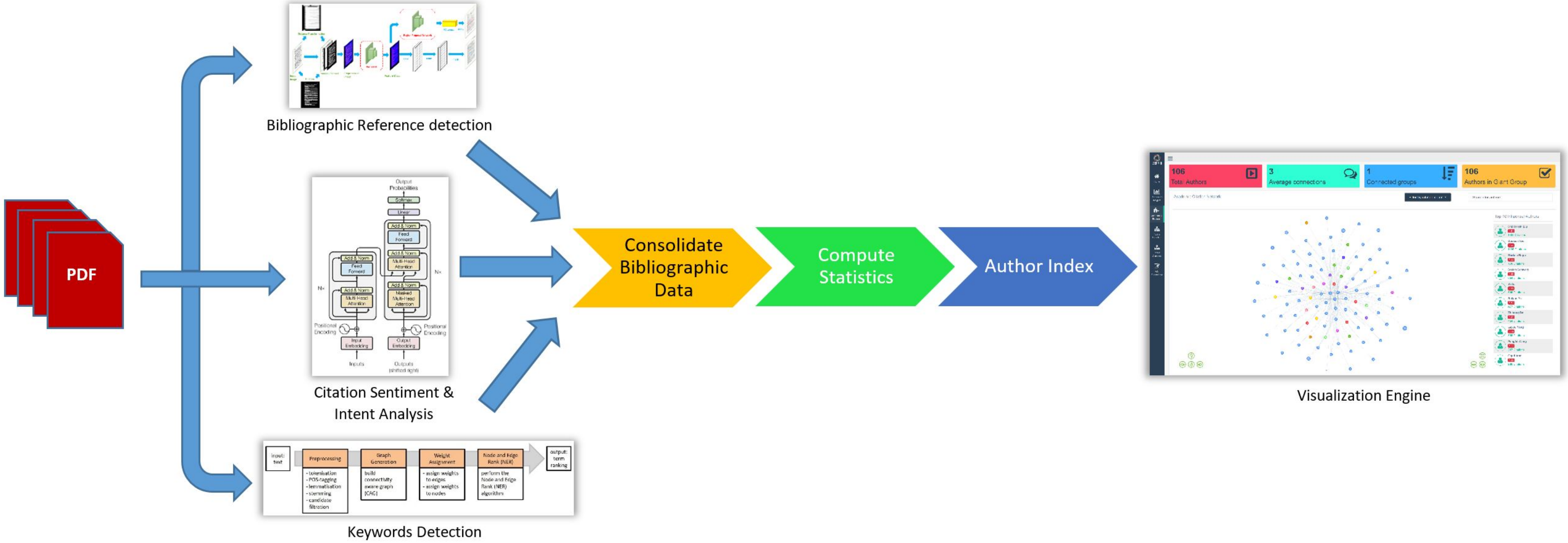


Figure 6.1: Overview of Academic Community Explorer 2.0 System

6.2.1 *Consolidate Bibliographic Data*

Data extracted by the Data Extraction modules in Stage 1 is consolidated into a common data storage. For this purpose, ACE employs MongoDB as the central storage where all the data is collected and secured. One of the challenges faced during the consolidation of data was to precisely identify each author and correctly assign the respective publications to the right author. This phenomenon is known as Author name disambiguation. One of the reasons which gave rise to this challenge is the use of abbreviated names in the reference section of the publications. For instance, there are two persons with names Anthony Davidson and Andrew Davidson. Both persons write their short name as A. Davidson. In a scenario where we only see the shortened name, it is very challenging to identify which specific person is being referred to in this name. Another challenge could appear if there is an error in the extracted text. The text from detected bibliographic references is extracted by performing Optical Character Recognition (OCR). There is a possibility for the introduction of OCR error in the extracted text. Especially, in the case of name initials, any misclassification can lead to an entirely different name for a person.

To tackle the challenges in author name disambiguation and ensure the quality of data, we employed a set of external resources i.e. Crossref and Semantic Scholar to validate the accuracy of extracted data. Both of these external resources have a huge collection of bibliographic data and have specified interfaces through which we can query data about a specific publication. Any error or disambiguation in names that arise during the data extraction phase is eradicated by verifying it with author names collected from the external resources. The consolidated data is now ready for further processing,

6.2.2 *Computation of Statistics*

This module deals with analyzing the consolidated data for estimation of certain author level and community level statistics. These statistics are crucial in identifying the author and community-level trends in different aspects.

Although, there are numerous fine-grained statistical figures extracted from the consolidated data i.e. number of publications or citations for an author, etc. However, we will only discuss the most prominent statistical figures estimated in this module. One of such figures is the generation of co-authors graph. We represented all authors as nodes such that all co-authors have a unidirectional link with one another. Once the co-authorship graph for a whole community is ready, then we apply Girwan Newmann clustering [47] on the co-authorship graph. It results in clusters of co-authors.

On the other hand, we also generate a community network graph. Given the consolidated data, we take authors in a community and represent them as nodes in a community network graph. Additionally, we employed the citation data to draw links between network nodes. So the resultant graph is a citation network graph. To incorporate more information in the citation network graph we color-coded the nodes based on their co-authorship cluster.

6.2.3 *Author Semantic Index*

This is an important module of stage 2, as it estimates the significance and influence of researchers in a scientific community which is one of the core features of ACE 2.0 System. For this purpose, we employ our proposed Semantic index. The Semantic Index has already been discussed in detail in Chapter 5. Estimation of index value concludes the data processing and analysis in stage 2.

6.3 STAGE3: VISUALIZATION ENGINE

Once all the statistics have been successfully estimated, the final data is delivered to the visualization engine, which uses the given data and visualizes in more than one way to highlight different trends. There are different visualizations with a granularity that spans over three levels. The highest level contains the visualization representing the domain-level insights, followed by less detailed visualization representing the community-level insights, and finally the author-level insights. The visualizations for each of these granularity levels are described below:

6.3.1 *Domain-Level Insights*

As the name suggests, these visualizations represent insights from the domain level. For the proof of concept, we selected the domain of Document Analysis as our sample domain. Fig. 6.2 shows overall statistics of the Document Analysis domain. The visualization shows three communities in the selected domain which have a total of 6255 authors who contributed 4638 publications with 65775 incoming and outgoing citations. However, the graph shown below the statistics depicts the interaction between communities. Each node in the visualization represents a community in the domain and the size of each node represents the number of citations it received. If the publications in a community receive relatively more citations, then their size will be bigger than other communities. The links between nodes represent the citation relation between two communities. The direction of the link represents the citation direction and the number on the link shows the number of times the publications of the target community

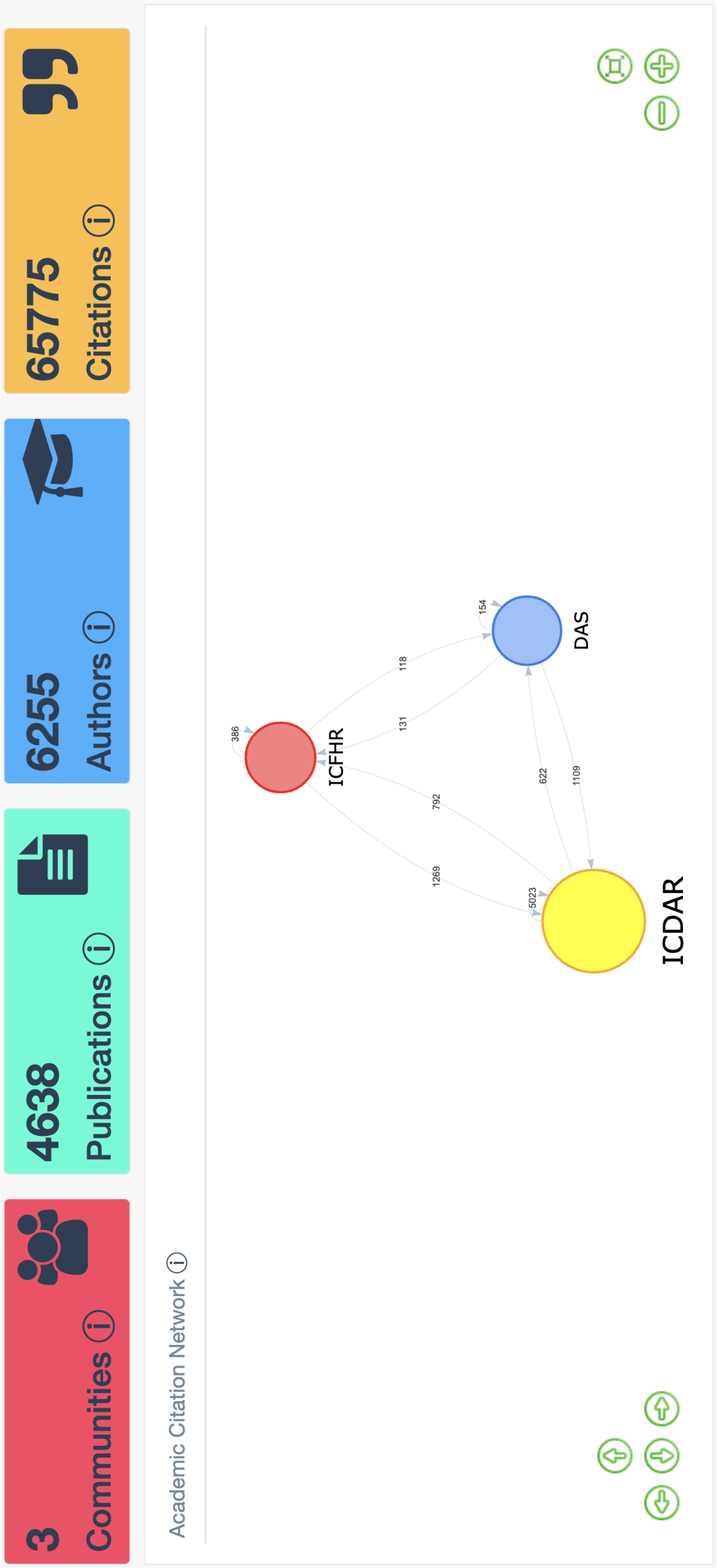
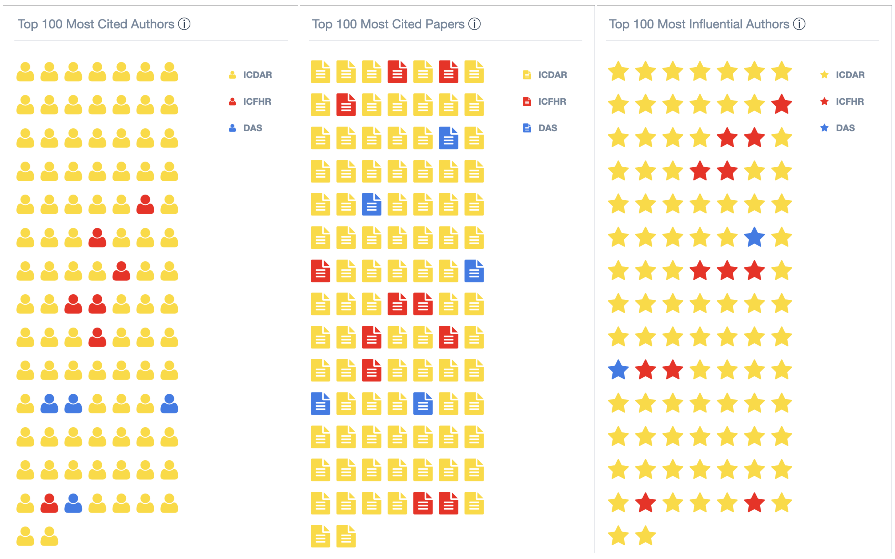


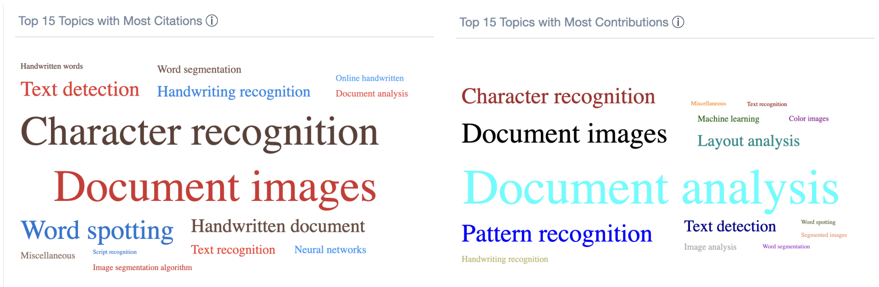
Figure 6.2: Community Interactions in the domain of Document Analysis



Top 100 Most Cited Authors ⓘ
ICDAR
ICFHR
DAS
Top 100 Most Cited Papers ⓘ
ICDAR
ICFHR
DAS
Top 100 Most Influential Authors ⓘ
ICDAR
ICFHR
DAS

(a) *Top Cited Authors* (b) *Top Cited Papers* (c) *Top Influential Authors*

Figure 6.3: Different Interaction patterns of an Author in a Scientific Community



(a) *Top Topics by Citations* (b) *Top Topics by Contributions*

Figure 6.4: Topic popularity in the Domain of Document Analysis

were cited by the other. Self-citations are also shown in the graph which is a key indicator to understanding the popularity of the publications within the community as well as in the other communities.

The next visualization in Fig. 6.3 shows the distribution of top publications and authors among all three communities. Fig. 6.3a and 6.3b show the distribution of top 100 authors and publications respectively in all communities ranked in the order of the number of citations received. However, Fig. 6.3a shows the distribution of the top 100 authors with the highest semantic index score in all communities. Lastly, Fig. 6.4a and 6.4b show the top topics with most citations and contributions among all communities. The size of the topic refers to the number of citations or contributions received by that topic.

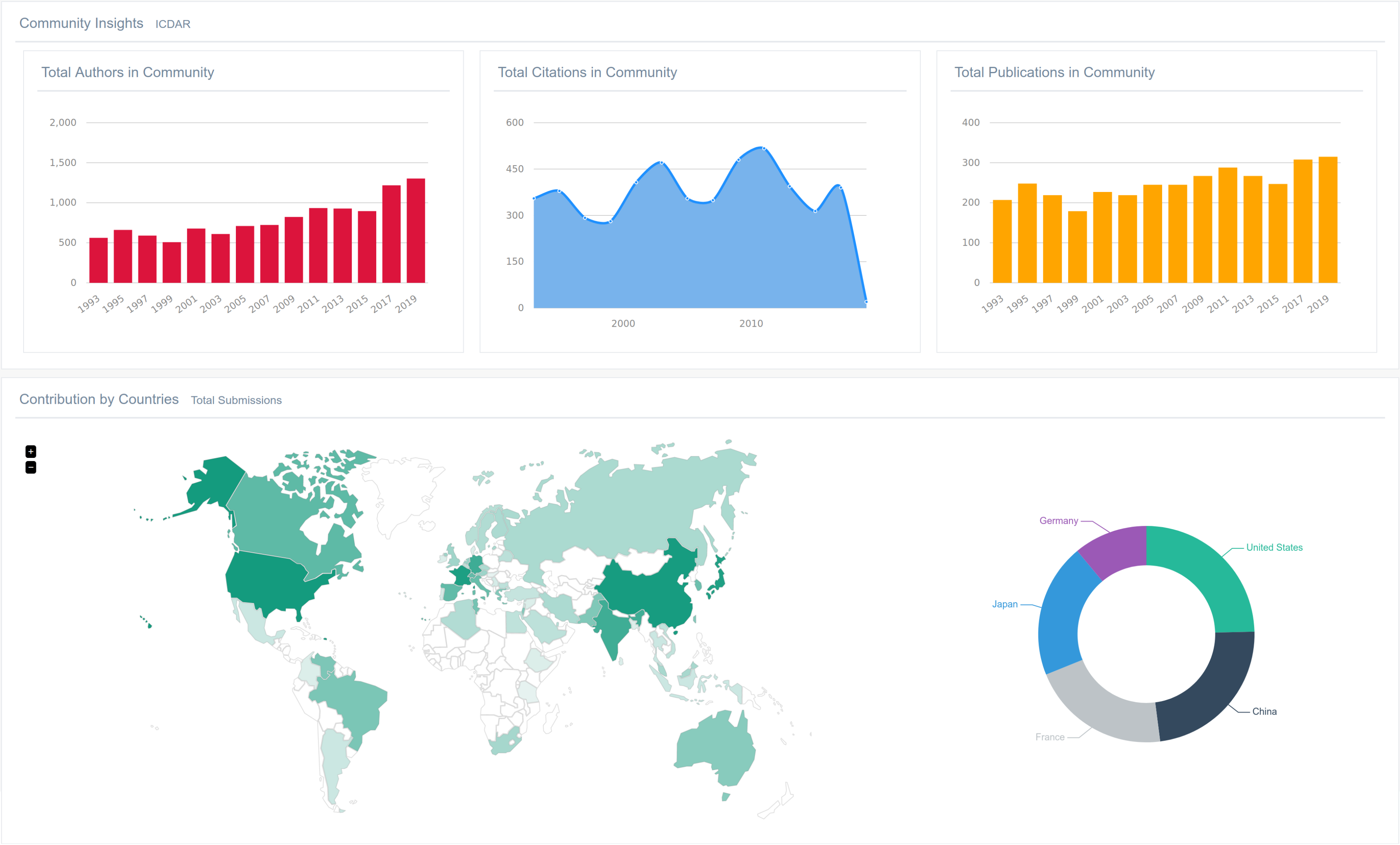


Figure 6.5: Example of overall community highlights

6.3.2 *Community-Level Insights*

This section describes the community-level insights which are more abstract than author-level highlights, however, they are more precise about the overall community. Such highlights play an important role in policing and paving the future path of a scientific community.

6.3.2.1 *Community Highlights*

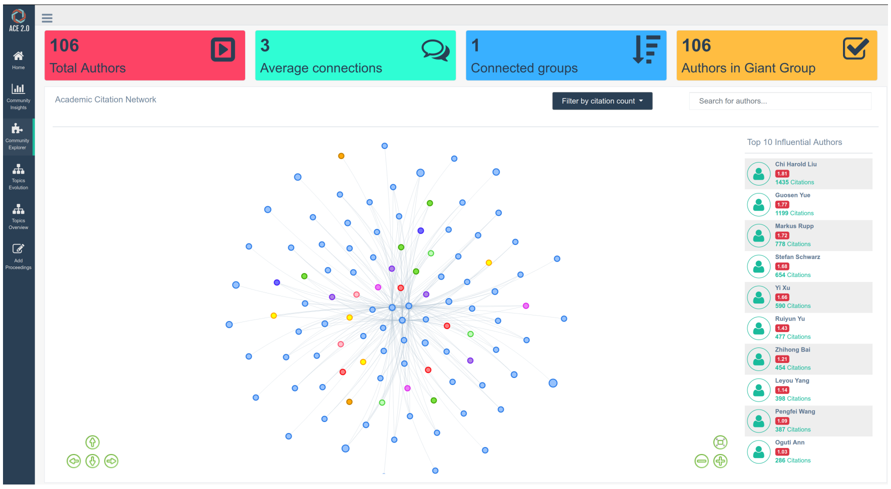
Overall community highlights give us a quick insight into community-level statistics. Fig 6.5 shows an example of different visualizations related to a community. A red bar chart shows the number of authors who participated in a proceeding year. It allows us to gauge if the number of participants is increasing or decreasing over time. On the other hand, the number of submissions received each year is represented in the orange bar chart. We also visualize the number of citations received by a publication by each proceeding year with the blue chart. It provides an immediate insight into which years are more popular within the community. From the example given, it is clear that the proceedings of the year 2011 are the most popular and have received the highest number of citations. Furthermore, we also visualize participation by country to see which countries are contributing the most to this academic community. It can be observed that the United States has the most contributions among all countries in the given scientific community.

6.3.2.2 *Community as a Network*

This section describes an important community-level feature highlight where the whole community is represented in the form of a community network graph. Fig 6.6 shows the example of an academic community, where each node represents an author. Authors are connected with links among them. Each link between two nodes represents a citation relationship. The color of the nodes represents the collaboration groups in the scientific community. Author nodes with the same color tend to collaborate in their research work. The network graph can be filtered using the citation count threshold. It will filter the graph and only shows the authors having the specified number of overall citations or more.

6.3.2.3 *Topic Evolution & its Trends*

This section describes the visualization related to topic evolution. The Fig 6.7 shows a dynamic visualization of the evolution of the topics over a period of time. It represents topics in the form of bubbles. The year slider on the top middle of the chart can be moved to see the effect on individual topic bubbles. When you move the slider, a topic bubble might become larger or smaller, representing its popularity



106 Total Authors

3 Average connections

1 Connected groups

106 Authors in Giant Group

Academic Citation Network

Top 10 Influential Authors

Chi Harold Liu	1.81	1435 Citations
Guosen Yue	1.77	1199 Citations
Markus Rupp	1.72	778 Citations
Stefan Schwarz	1.68	654 Citations
Yi Xu	1.66	590 Citations
Ruiyun Yu	1.43	477 Citations
Zhihong Bai	1.21	454 Citations
Leyou Yang	1.14	398 Citations
Pengfei Wang	1.09	387 Citations
Oguti Ann	1.03	286 Citations

Figure 6.6: Community as a Network

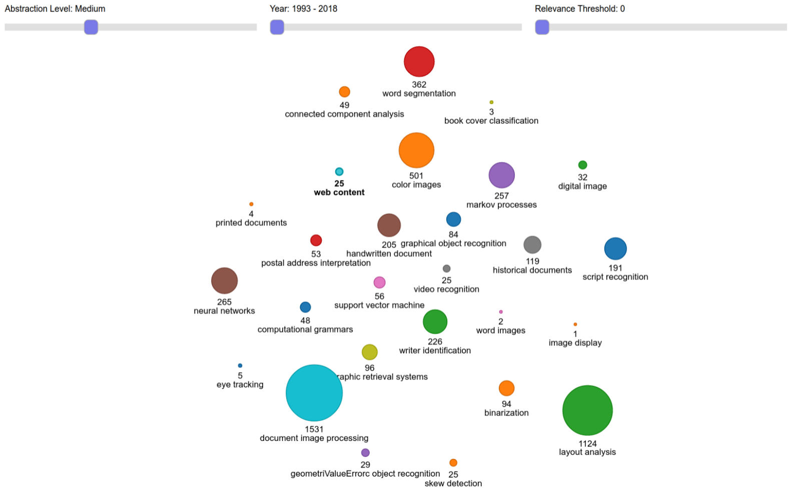


Figure 6.7: Dynamic representation of Topic Evolution

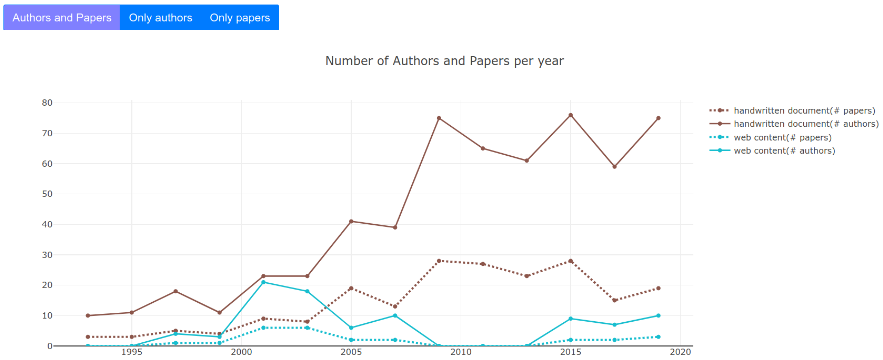
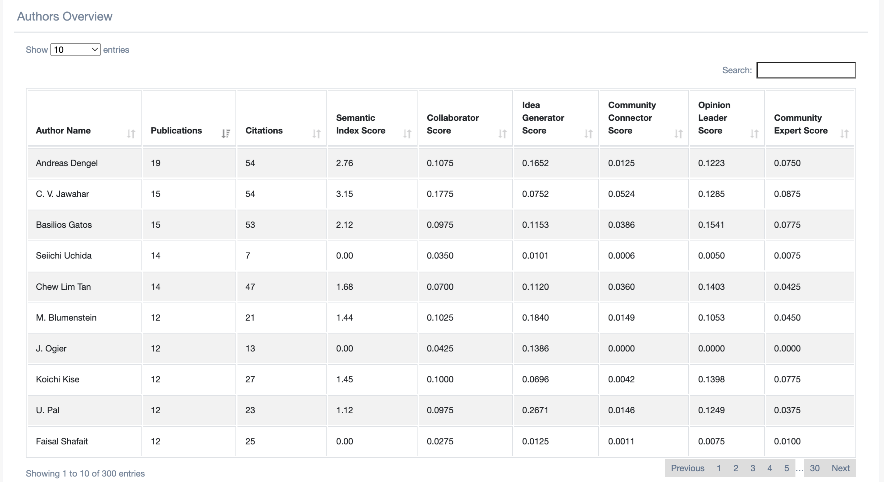


Figure 6.8: Visualizing Topic evolution of selected Topics



Authors Overview

Show 10 entries

Search:

Author Name	Publications	Citations	Semantic Index Score	Collaborator Score	Idea Generator Score	Community Connector Score	Opinion Leader Score	Community Expert Score
Andreas Dengel	19	54	2.76	0.1075	0.1652	0.0125	0.1223	0.0750
C. V. Jawahar	15	54	3.15	0.1775	0.0752	0.0524	0.1285	0.0875
Basilios Gatos	15	53	2.12	0.0975	0.1153	0.0386	0.1541	0.0775
Seiichi Uchida	14	7	0.00	0.0350	0.0101	0.0006	0.0050	0.0075
Chew Lim Tan	14	47	1.68	0.0700	0.1120	0.0360	0.1403	0.0425
M. Blumenstein	12	21	1.44	0.1025	0.1840	0.0149	0.1053	0.0450
J. Ogier	12	13	0.00	0.0425	0.1386	0.0000	0.0000	0.0000
Koichi Kise	12	27	1.45	0.1000	0.0696	0.0042	0.1398	0.0775
U. Pal	12	23	1.12	0.0975	0.2671	0.0146	0.1249	0.0375
Faisal Shafait	12	25	0.00	0.0275	0.0125	0.0011	0.0075	0.0100

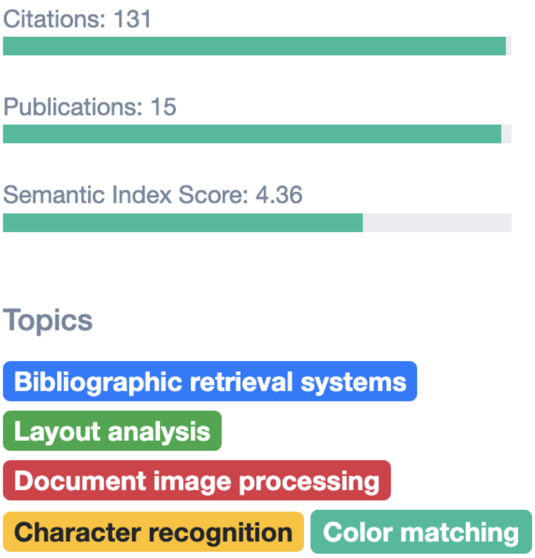
Showing 1 to 10 of 300 entries

Figure 6.9: Authors Overview

in the selected year. Fig 6.8 shows an example of two selected topics along with the number of contributions for these topics over the years. The topic "Handwritten Document" and "Web Content" are represented in brown and blue colors respectively. Solid lines represent the number of scientific publications submitted on a specific topic. On the other hand, dotted lines represent the number of authors who contributed to a specific topic. It is quite evident that both topics started with minimal interest at the start. With time, the topic of 'Handwritten Documents' became increasingly popular within the scientific community. Contrary to this, the topic of "Web Content" gradually increased in popularity. However, after the year 2007 scientific community lost interest in this topic. Such trends help us understand the community's interests and make decisions regarding the future direction of the community.

6.3.2.4 *Authors Overview*

The table shown in Fig. 6.9 shows different statistics for all authors in a scientific community. These statistics range from simple citation



Citations: 131

Publications: 15

Semantic Index Score: 4.36

Topics

Bibliographic retrieval systems

Layout analysis

Document image processing

Character recognition

Color matching

Figure 6.10: Author Statistics Overview

or publication count to complex semantic index values. The authors can be sorted with respect to any measure in the table by clicking the title of the measure of interest. It provides a quick quantitative comparison between different authors in a single glimpse.

6.3.3 *Author-Level Insights*

This section discusses the visualizations designed for displaying the author-level statistics. Some key author-level visualizations are as follows:

6.3.3.1 *Author Statistics*

Fig. 6.10 shows how the basic statistics of a specific author are displayed in the author's profile. The bars on the top show the citation count, the number of publications, and the semantic index score of a given author. The extent to which the bars are filled represents their percentile. On the bottom, we can see the top 5 topics on which this author is continuing their research.

6.3.3.2 *Community Roles*

This section discusses the roles of an author in an academic community. In this work, we consider each author for five different roles. These roles are represented by different centrality measures discussed in Section 5.3. Different roles and their description are as follows:

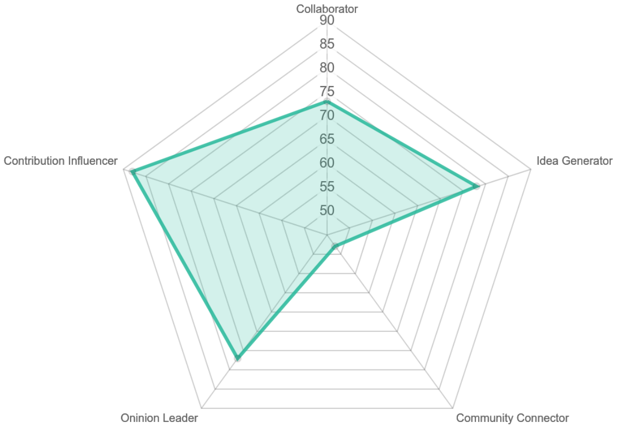


Figure 6.11: Visualization of an Author's Roles in a Community

- **Collaborator:** Collaborates more often with members of the community. It is represented by the Degree centrality of the author.
- **Idea Generator:** Highly influential individual, who brings new ideas which are widely accepted by the community. It is represented by the Eigenvector centrality of the author.
- **Community Connector:** Diversely publishes with different cliques in the community. It is represented by the Betweenness centrality of the author.
- **Opinion Leader:** Holds a strong network and is capable of influencing an opinion about a trend in the community. It is represented by the Closeness centrality of the author.
- **Contribution Influencer:** Dominates the community with their important scientific contributions. It is represented by the Indegree Centrality of the author.

Fig 6.11 represents a visualization example of an author's role in the community. It can be noticed that in this specific example the person in the discussion is more of a contribution influencer as compared to any other role in the community. Fig 6.12 shows a set of sliders for each role that can slide to increase or decrease the extent to which they contribute towards estimating the semantic index.

6.3.3.3 *Citation Sentiment Analysis*

In this section, we will discuss the system's features related to the citation sentiment of an author in a scientific community. Fig 6.13 shows an example visualization for the overall citation sentiment of an author. A doughnut chart is used for this visualization. Positive, negative, and neutral citation sentiments are represented in green, red, and gray colors. It can be seen that this specific author has mostly

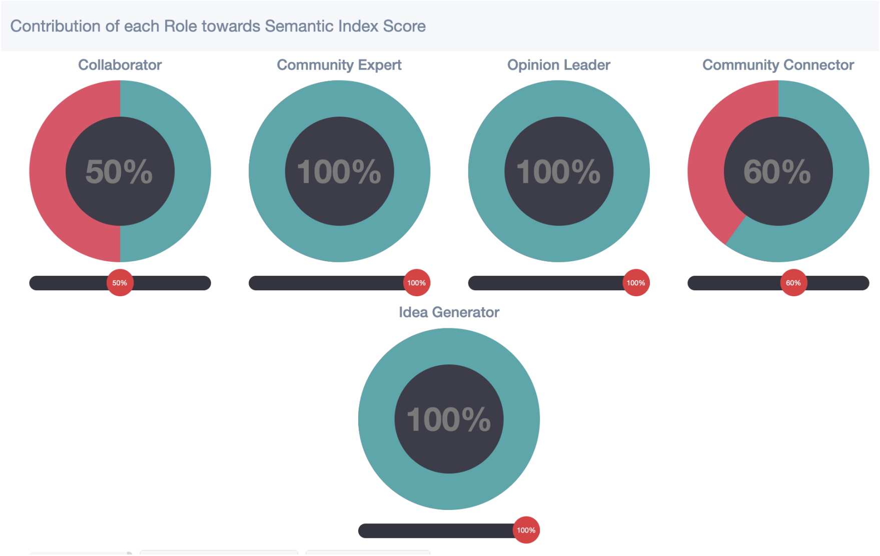


Figure 6.12: Roles Contribution Sliders

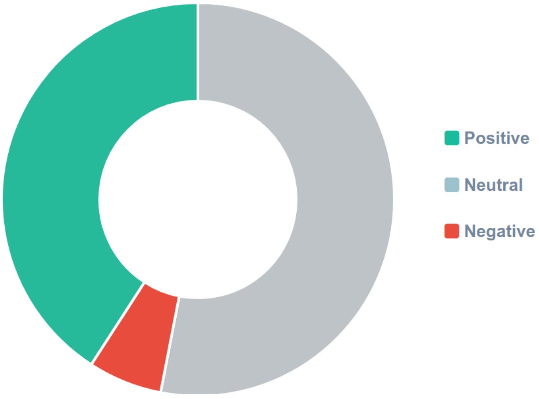
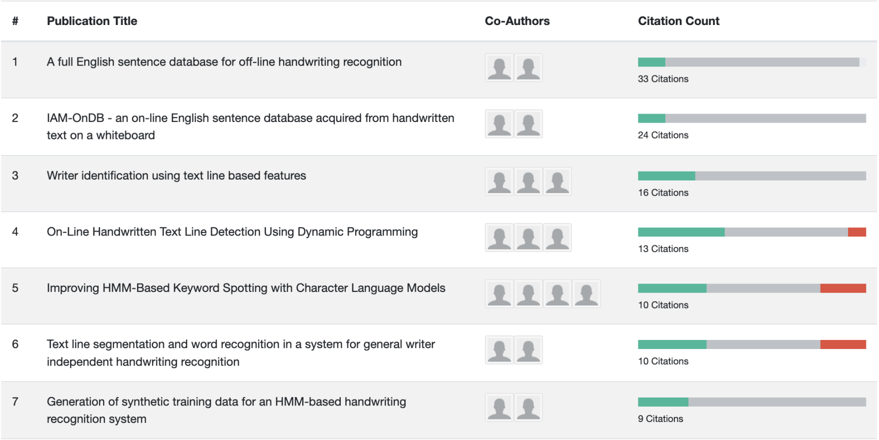
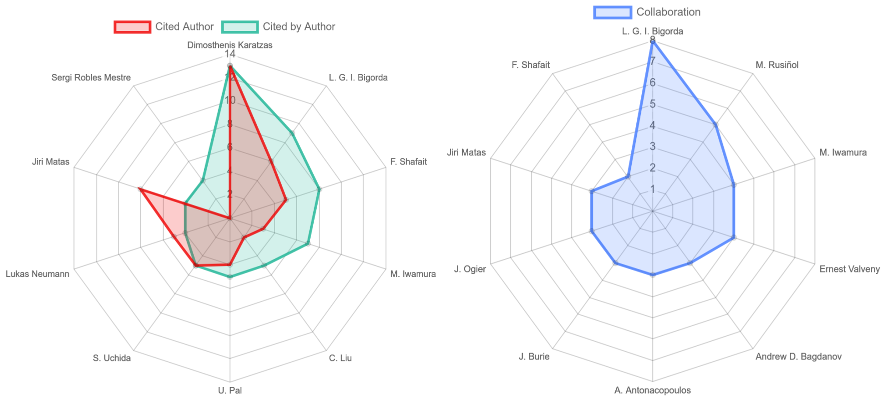


Figure 6.13: Visualization of Citation Sentiment all the Author's citations in a Community



#	Publication Title	Co-Authors	Citation Count
1	A full English sentence database for off-line handwriting recognition		33 Citations
2	IAM-OnDB - an on-line English sentence database acquired from handwritten text on a whiteboard		24 Citations
3	Writer identification using text line based features		16 Citations
4	On-Line Handwritten Text Line Detection Using Dynamic Programming		13 Citations
5	Improving HMM-Based Keyword Spotting with Character Language Models		10 Citations
6	Text line segmentation and word recognition in a system for general writer independent handwriting recognition		10 Citations
7	Generation of synthetic training data for an HMM-based handwriting recognition system		9 Citations

Figure 6.14: Publications List with Sentiment



(a) *Citation Pattern*

(b) *Collaboration Patter*

Figure 6.15: Different Interaction patterns of an Author in a Scientific Community

received neutral citations. However, positive citations also have a fair share in total citations which shows that the scientific contributions by the author in the discussion have been fairly accepted and appreciated in the academic community. Fig. 6.14 shows a list of publications of a given author. The last column on the right shows the citation sentiment of all the citations received by each publication.

6.3.3.4 *Author Citation Patterns*

This section discusses the citation pattern of individual authors. The Fig 6.15a shows a visualization example of a radar chart representing the citation pattern of an author. The visualization shows the top 10 authors who were cited by the author in the discussion and the number of times they were cited is represented in green color. However, the data points in red color represent how many times those authors cited back the author in the discussion. With this visualization, we can instantly realize the citation interaction of an author with the other community members. In the given example we can see that the author in the discussion cited himself more than any other else in the community.

6.3.3.5 *Author Collaboration*

This section presents the collaboration pattern of an author. Fig 6.15b shows an example collaboration visualization of an author. Each data point shows the number of times this author collaborated with other researchers in the academic community and is represented in blue color. This visualization in combination with citation pattern visualization can uncover even more patterns. For example, in the given examples, we can see that the current author tends not to cite his second most common collaborator very often.

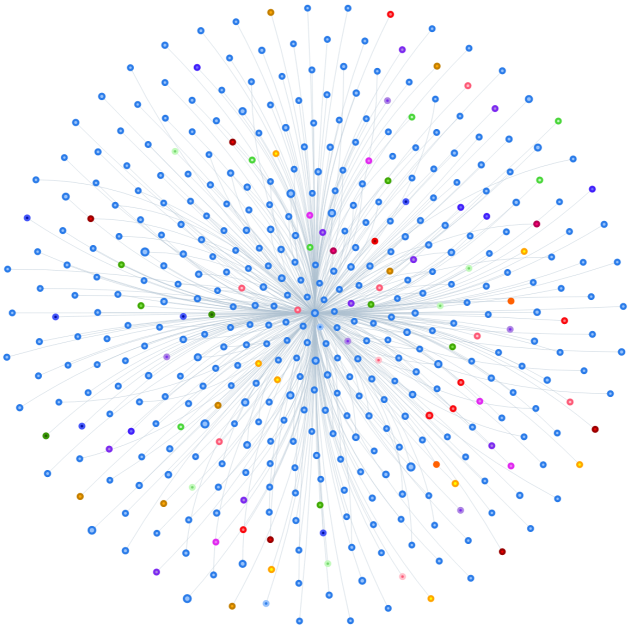


Figure 6.16: Visualization of Author's custom Network in a Community

6.3.3.6 *Customized Author Network*

Our proposed tool also includes a customized network graph on each author's profile page. Fig 6.16 shows an example of a custom network of an author. Where the author in the discussion is at the center of the network graph. Each node represents other authors in the academic community who either cited or were cited by the current author. With this visualization, we can get a quick insight into the extent of networking of an author in the academic community.

Part V

CONCLUSION

7 CONCLUSION & FUTURE WORK

In this chapter, we will review the challenges we faced during the course of this research, followed by the solutions we have proposed in this thesis. This thesis consists of three main components: data extraction, analysis, and visualization. We will discuss the challenges faced by each of these components, as well as the respective proposed solutions.

With the growing number of publications every year, it is increasingly difficult to evaluate the quality of publications without relying on automatic methods. It is imperative for libraries to index all publications with their references. The first component of this thesis is data extraction. To be precise, we are interested in extracting bibliographic references from scientific publications. Existing approaches rely heavily on textual features to detect and extract bibliographic references. Usually, text-based approaches are very effective. It is more difficult to achieve generalizability with different types of existing referencing styles, Text-based approaches find it very challenging to detect a referencing style that was not part of the training set. To resolve this issue, we proposed a layout-based approach that uses layout features to identify and extract bibliographic references from scientific publications. Our solution relies on layout features, which enables it to be independent of domain, language, and referencing style. As a result, the goal of generalizability is achieved. Furthermore, we also released the largest publicly available dataset for reference detection using layout features. The dataset contains manually annotated images containing thousands of references.

Keywords provide a meaningful understanding and central idea of a given text. Scientific publications also include keywords. However, these keywords do not necessarily reflect the actual topic of the publication. Keywords are sometimes included in a publication solely to comply with the topic requirements of a specific publication venue. Therefore, the mentioned topics are not reliable. For this purpose, we introduced a two-stage system called CoCoNoW. Its first stage is responsible for extracting meaningful keywords. CoCoNoW builds a connectivity-aware graph of the terms and assigns them meaningful weights. It then sorts the terms in order of relevance. In the second stage, those keywords along with an ontology are used to identify the topics of the publication. A domain ontology serves as an attention-map/context for topic modeling based on the detected keywords. An ontology makes it possible to generalize and adapt the system to any domain. We evaluated CoCoNoW on 3 public datasets, and it achieved state-of-the-art performance on each of those datasets.

Sentiment analysis is a popular task. The majority of the existing literature focuses on Twitter/product reviews sentiment analysis. There has been limited research performed to analyze the sentiment of citations. Citations are an essential part of measuring the impact of a researcher or publication within a scientific community. Usually, citations are considered to be a quantitative measure. However, sentiment and intent analysis of a citation provide us with the qualitative aspect of a citation. Therefore, citation impact analysis enables us to quantify the quality of a citation. Existing approaches analyze the sentiment of a citation. However, their training is highly biased due to the highly imbalanced class distribution. It is a common observation that the neutral class has a significantly higher number of samples than either the positive or negative classes. Hence, the model trained on such datasets is likely to show a bias towards the neutral class. In order to tackle this problem, we proposed ImpactCite an XLNet-based approach. ImpactCite was evaluated on publicly available datasets and demonstrated state-of-the-art performance. It is worth noting that traditional models have a single model for each task. However, the ImpactCite is a multitask model. This means that a single model can classify both the Sentiment and intent of a given citation string. Another notable challenge for citation impact analysis is the scarcity of data and the associated costs of data annotation. To address this issue, we explored the impact of using an out-of-domain data set to pre-train our model before fine-tuning it on our target domain dataset. We explored different data feeding strategies. According to the evaluation results, sequential data scheduling is more suited for domain-specific use cases. However, feeding shuffled data results in a more generalized model which is suitable for a generic use case. Additionally, we released a cleaned version of a publicly available dataset called Citation Sentiment Corpus. We removed all the inconsistent and duplicate samples in different sets. The new clean CSC dataset is therefore a better and more reliable dataset for the task of citation sentiment analysis.

Influential personalities have always been a part of any society. In scientific communities, the influence of a person is generally measured by its citation count. In other words, raw citation count serves as a tool for measuring a person's influence in a scientific community. The scientific community introduced various author indices to assess the quality of a research artifact, which relied solely on the raw citation count. With the passage of time, the mechanism for defining those author indexes became increasingly sophisticated. Several mechanisms were developed to filter the number of citations that are to be counted when estimating the author index. Although each of the existing author indexes was carefully designed, the existing author indexes each have certain limitations. Some indexes are biased in favor of the older researchers and to the detriment of the newer researchers. Some indexes are misleading or unsuitable for assessing the quality of research work. After analyzing citation impact in this thesis, we found that the meaning of a citation can change significantly depending on the purpose and circumstances and not all citations are of equal importance. Existing author indexes only take into account the quantitative aspect of the citations while completely overlooking the qualitative aspect. In this thesis, we introduced a novel author index called the Semantic index, which takes into consideration both qualitative and quantitative aspects of citations. Additionally, the Semantic Index analyzes each researcher in the community for five different roles based on centrality measures. Where each centrality measure represents a specific role. i.e., idea generator, opinion leader, etc. The Semantic index also takes into account these centrality measures to determine the index value for a given researcher. The Semantic Index was evaluated according to the guidelines defined by DORA. These guidelines have been developed through a collaborative effort of dozens of organizations around the world with the aim of defining a meaningful way to assess the quality of research. The Semantic Index adheres to all the DORA guidelines, which makes it a more suitable alternative to the existing author indexes.

In today's digital era, it is remarkably convenient for researchers to share and collaborate on unique scientific ideas. Depending on the domain, people often strive to achieve these endeavors through tightly-knit scientific communities. Technological advancements and their evolution over time have spurred the emergence of research communities with unique topics and focus. It is a challenging task to police scientific communities and administer them from a quantitative and qualitative perspective due to the enormous number of scientific communities and their vastness. Existing approaches provide limited and shallow insight into a scientific community. We proposed ACE 2.0, a tool for scientific communities, which employs state-of-the-art models to automatically, efficiently, and smartly extract, and analyze scientific data. Furthermore, it provides a wide range of insights that cover individual researchers to the entire community. These insights include different community-level aspects, such as collaboration patterns, citation patterns, influential persons with different roles, contributions from geographical locations, topic evolution, and many other fine-grained aspects within each scientific community. Our system considers scientific publications to be the primary source of information, but it also draws on several external resources to collect as much data as possible in order to correctly identify individual researchers and their contributions. ACE 2.0 performs an analysis of scientific communities and automatically performs detailed digital profiling of individual researchers. It analyzes their information to identify trends in citation, collaboration, contributions, popularity, and their role in the community. ACE 2.0 also proposes a new index for contributing researchers in the scientific community, which takes into account both quantitative and qualitative aspects of an individual citation. This work motivates us to discover endless new perspectives and opens it to a wide range of applications in other domains.

7.1 LIMITATIONS

This thesis proposed a comprehensive system called ACE for getting insights into a scientific community. For this purpose, ACE needs all the data for a selected domain. It expects the publications as input in the form of born-digital Portable Document Format (PDF)s. However, the availability of complete conference proceedings is the biggest challenge, as the publishers own the right to provide the proceedings to the public. It is a common occurrence that those proceedings are made available to premium subscribers only, who pay a subscription fee to access those proceedings. Additionally, there are multiple publishers with each having their own subscriptions, which makes it impractical to obtain all the proceedings from a given domain.

Secondly, ACE relies on external resources to perform author name disambiguation. In general, it works very well for some domains. However, the literature coverage of the external resources plays a crucial role in defining the robustness. For instance, the literature coverage of external resources for the social science domain has limited to no coverage. Sometimes, the relevant records are found, however, they have missing attributes. This makes it more challenging for us to perform author name disambiguation for all possible domains.

Additionally, ACE represents authors as nodes and uses colors to encode the co-authorship relation between authors, where each cluster of co-authors is assigned the same color. The limitation arises when more and more data is processed and the number of clusters increases exponentially. With this increase in the number of clusters, it becomes infeasible for the visualization to assign visually different colors. At times, the colors assigned to two different nodes are not visibly differentiable.

7.2 FUTURE WORK

The work proposed in this thesis is a proof of concept that we can use the research publications of scientific communities to get insights into scientific communities. For this thesis, we selected the domain of Document Analysis along with three communities within this domain. The potential of this work could be fully explored by processing an even larger amount of data from a wider variety of domains, for example, computer vision, which has many more communities and which has communities that are significantly larger. Even some communities have tens of thousands of publications. Another approach would be to study the interaction between different domains by examining data from multiple domains. For example, bioinformatics links together two different scientific fields.

Another important future study for this thesis would be to examine the shift in research direction by a researcher. With the constantly evolving field of research, it is interesting to know if several researchers from a specific topic are switching to another newly evolving topic. Such trends demonstrate the dynamic shift in interest within the community.

This thesis demonstrated that the system already works for established researchers. Furthermore, it can be adjusted to highlight both existing and rising talent in the research community. This thesis work could be developed into a Talent Recommender System. The objective is to gather as much data as possible for a specific domain, along with all the researchers associated with it. After evaluating the impact of each individual's research, we compile a database of all notable or rising stars of the community. Upon searching for a specific topic in that domain, the recommender system can provide us with existing or upcoming talent in the community. A recommender system of this nature would be of great value to funding agencies, universities, and talent scouts.

BIBLIOGRAPHY

[1] D. Mercier*, S. T. R. Rizvi*, V. Rajashekar, A. Dengel, and S. Ahmed. “ImpactCite: An XLNet-based solution enabling qualitative citation impact analysis utilizing sentiment and intent.” In: *ICAART 2021 - Proceedings of the 13th International Conference on Agents and Artificial Intelligence* 2 (2021), pp. 159–168. DOI: 10.5220/0010235201590168 (cit. on pp. 61, 62, 66–69, 76, 77).

[2] A. Abu-Jbara, J. Ezra, and D. Radev. “Purpose and Polarity of Citation: Towards NLP-based Bibliometrics.” In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 596–606. URL: https://www.aclweb.org/anthology/N13-1067 (cit. on p. 56).

[3] M. W. Ahmed and M. T. Afzal. “FLAG-PDFe: Features Oriented Metadata Extraction Framework for Scientific Publications.” In: *IEEE Access* 8 (2020), pp. 99458–99469 (cit. on p. 15).

[4] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes. “Sensing trending topics in Twitter.” In: *IEEE Transactions on Multimedia* 15.6 (2013), pp. 1268–1282 (cit. on p. 42).

[5] S. Alonso, F. Cabrerizo, E. Herrera-Viedma, and F. Herrera. “hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices.” In: *Scientometrics* 82.2 (2010), pp. 391–400 (cit. on pp. 79, 82).

[6] *AnyStyle*. https://anystyle.io/ (cit. on p. 14).

[7] A. R. Aronson, O. Bodenreider, H. F. Chang, S. M. Humphrey, J. G. Mork, S. J. Nelson, T. C. Rindflesch, and W. J. Wilbur. “The NLM Indexing Initiative.” In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association. 2000, p. 17 (cit. on pp. 45–47).

[8] A. Athar. “Sentiment Analysis of Citations using Sentence Structure-Based Features.” In: *Proceedings of the ACL 2011 Student Session*. Portland, OR, USA: Association for Computational Linguistics, June 2011, pp. 81–87. URL: https://www.aclweb.org/anthology/P11-3015 (cit. on pp. 55, 59, 68, 69).

[9] S.-A. Bahrainian and A. Dengel. “Sentiment analysis and summarization of twitter data.” In: *2013 IEEE 16th International Conference on Computational Science and Engineering*. IEEE. 2013, pp. 227–234 (cit. on p. 53).

[10] H. Becker, M. Naaman, and L. Gravano. "Beyond trending topics: Real world event identification on twitter." In: *AAAI*. 2011 (cit. on p. 42).

[11] S. Beliga. "Keyword extraction: a review of methods and approaches." In: *University of Rijeka, Department of Informatics* (2014), pp. 1–9 (cit. on p. 37).

[12] A. Bellaachia and M. Al-Dhelaan. "Ne-rank: A novel graph-based keyphrase extraction in twitter." In: *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Vol. 1. IEEE. 2012, pp. 372–379 (cit. on p. 41).

[13] I. Beltagy, K. Lo, and A. Cohan. "SciBERT: A pretrained language model for scientific text." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 3606–3611 (cit. on pp. 54, 56, 67).

[14] A. Bhardwaj, L. Erhard, A. Klein, S. Zander, and P. Zumstein. *ICONIP Datasaet: Labeled Reference Data from the Linked Open Citation Database (LOC-DB) Project*. `https://madata.bib.uni-mannheim.de/id/eprint/268`. Feb. 2018. DOI: `https://doi.org/10.7801/268` (cit. on pp. 16, 23).

[15] A. Bhardwaj, D. Mercier, A. Dengel, and S. Ahmed. ""Deep-BIBX: Deep Learning for Image Based Bibliographic Data Extraction"." In: *"Neural Information Processing"*. Cham: Springer International Publishing, 2017, pp. 286–293 (cit. on pp. 16, 17, 22–24, 26, 27).

[16] *Biblio*. `https://metacpan.org/release/MJEWELL/Biblio-Citation-Parser-1.10` (cit. on p. 14).

[17] S. K. Biswas, M. Bordoloi, and J. Shreya. "A graph based keyword extraction model using collective node weight." In: *Expert Systems with Applications* 97 (2018), pp. 51–59 (cit. on pp. 37–40, 42).

[18] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute. "A Principal Component Analysis of 39 Scientific Impact Measures." In: *PLOS ONE* 4.6 (June 2009), pp. 1–11. DOI: `10.1371/journal.pone.0006022` (cit. on p. 79).

[19] F. Boudin. "Unsupervised keyphrase extraction with multipartite graphs." In: *arXiv preprint arXiv:1803.08721* (2018) (cit. on p. 37).

[20] Z. Boukhers, S. Ambhore, and S. Staab. "An End-to-End Approach for Extracting and Segmenting High-Variance References from PDF Documents." In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 2019, pp. 186–195 (cit. on p. 15).

[21] T. D. Breaux and J. W. Reed. "Using ontology in hierarchical information clustering." In: *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE. 2005, 111b–111b (cit. on p. 43).

[22] S. Brin and L. Page. "The anatomy of a large-scale hypertextual web search engine." In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117 (cit. on p. 87).

[23] R. J. Brown. "A simple method for excluding self-citation from the h-index: the b-index." In: *Online Information Review* (2009) (cit. on p. 85).

[24] L. Cai, J. Tian, J. Liu, X. Bai, I. Lee, X. Kong, and F. Xia. "Scholarly impact assessment: a survey of citation weighting solutions." In: *Scientometrics* 118.2 (2019), pp. 453–478 (cit. on p. 79).

[25] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. Coronado, and J. Oliver. "Level statistics of words: Finding keywords in literary texts and symbolic sequences." In: *Physical Review E* 79.3 (2009), p. 035102 (cit. on p. 37).

[26] C. Carretero-Campos, P. Bernaola-Galván, A. Coronado, and P. Carpena. "Improving statistical keyword detection in short texts: Entropic and clustering approaches." In: *Physica A: Statistical Mechanics and its Applications* 392.6 (2013), pp. 1481–1492 (cit. on p. 37).

[27] R. Carston. "Linguistic communication and the semantics/pragmatics distinction." In: *Synthese* 165.3 (2008), pp. 321–345 (cit. on p. 38).

[28] C. Chen, K. Yang, H. Kao, and J. Ho. "BibPro: A Citation Parser Based on Sequence Alignment Techniques." In: *22nd International Conference on Advanced Information Networking and Applications - Workshops (aina workshops 2008)*. Mar. 2008, pp. 1175–1180. DOI: 10.1109/WAINA.2008.125 (cit. on p. 14).

[29] *Citaion Parser*. https://github.com/manishbisht/Citation-Parser (cit. on pp. 11, 14).

[30] M. Cliche. "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs." In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 573–580. DOI: 10.18653/v1/S17-2094 (cit. on p. 55).

[31] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady. "Structural scaffolds for citation intent classification in scientific publications." In: *arXiv preprint arXiv:1904.01608* (2019) (cit. on pp. 56, 61, 62, 65, 67).

[32] R. Costas and M. Bordons. "Is g-index better than h-index? An exploratory study at the individual level." In: *Scientometrics* 77.2 (2008), pp. 267–288 (cit. on pp. 79, 81).

[33] I. G. Councill, C. L. Giles, and M. Kan. "ParsCit: an Open-source CRF Reference String Parsing Package." In: *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. 2008 (cit. on pp. 15, 26, 28–31).

[34] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. "Transformer-xl: Attentive language models beyond a fixed-length context." In: *arXiv preprint arXiv:1901.02860* (2019) (cit. on p. 63).

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on pp. 55, 65–69, 76).

[36] S. Duari and V. Bhatnagar. "sCAKE: Semantic Connectivity Aware Keyword Extraction." In: *Information Sciences* 477 (2019), pp. 100–117 (cit. on pp. 37–40, 45, 47, 48).

[37] L. Egghe et al. "An improvement of the h-index: The g-index." In: *ISSI newsletter* 2.1 (2006), pp. 8–9 (cit. on p. 81).

[38] L. Erhard, A. Klein, S. T. R. Rizvi, S. Zander, and P. Zumstein. *RefDet Dataset: Additional Labeled Reference Data from the Linked Open Citation Database (LOC-DB) Project*. https://madata.bib.uni-mannheim.de/id/eprint/283. Jan. 2019. DOI: https://doi.org/10.7801/283 (cit. on pp. 17, 22).

[39] A. Esuli and F. Sebastiani. "Determining term subjectivity and term orientation for opinion mining." In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006 (cit. on p. 55).

[40] R. Feldman. "Techniques and applications for sentiment analysis." In: *Communications of the ACM* 56.4 (2013), pp. 82–89 (cit. on p. 53).

[41] C. Florescu and C. Caragea. "A position-biased pagerank algorithm for keyphrase extraction." In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017 (cit. on pp. 37, 38, 40, 41, 47).

[42] M. Franceschet and G. Colavizza. "TimeRank: A dynamic approach to rate scholars using citations." In: *Journal of Informetrics* 11.4 (2017), pp. 1128–1141. DOI: https://doi.org/10.1016/j.joi.2017.09.003 (cit. on p. 88).

[43] F. Franceschini and D. Maisano. "Criticism on the hg-index." In: *Scientometrics* 86.2 (2011), pp. 339–346 (cit. on p. 82).

[44] M. Goldberg. *free_cite*. https://github.com/miriam/free{_}cite (cit. on p. 15).

[45] S. Galam. "Tailor based allocations for multiple authorship: a fractional gh-index." In: *Scientometrics* 89.1 (2011), pp. 365–379 (cit. on p. 86).

[46] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. *Detectron*. https://github.com/facebookresearch/detectron . 2018 (cit. on p. 22).

[47] M. Girvan and M. E. J. Newman. "Community structure in social and biological networks." In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826. DOI: 10.1073/pnas.122653799. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.122653799 (cit. on p. 101).

[48] Google. *Google Scolar*. URL: https://scholar.google.com/ (visited on 06/30/2022) (cit. on p. 4).

[49] M. Grennan and J. Beel. *Synthetic vs. Real Reference Strings for Citation Parsing, and the Importance of Re-training and Out-Of-Sample Data for Meaningful Evaluations: Experiments with GROBID, GIANT and Cora*. 2020. arXiv: 2004.10410 [cs.LG] (cit. on p. 16).

[50] N. T. Hagen. "Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis." In: *PLoS One* 3.12 (2008), e4021 (cit. on p. 86).

[51] N. T. Hagen. "Harmonic publication and citation counting: sharing authorship credit equitably–not equally, geometrically or arithmetically." In: *Scientometrics* 84.3 (2010), pp. 785–793 (cit. on p. 86).

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask R-CNN." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322 (cit. on pp. 21, 22, 24).

[53] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778 (cit. on pp. 21, 22).

[54] J. E. Hirsch. "An index to quantify an individual's scientific research output." In: *Proceedings of the National Academy of Sciences* 102.46 (2005), pp. 16569–16572. DOI: 10.1073/pnas.0507655102. eprint: https://www.pnas.org/content/102/46/16569.full.pdf (cit. on p. 3).

[55] J. E. Hirsch. "An index to quantify an individual's scientific research output." In: *Proceedings of the National academy of Sciences* 102.46 (2005), pp. 16569–16572 (cit. on pp. 79, 80).

[56] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory." In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735 (cit. on p. 16).

[57] A. Hulth. "Improved automatic keyword extraction given more linguistic knowledge." In: *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics. 2003, pp. 216–223 (cit. on pp. 45–47).

[58] N. P. Hummon and P. Dereian. "Connectivity in a citation network: The development of DNA theory." In: *Social Networks* 11.1 (1989), pp. 39–63. DOI: https://doi.org/10.1016/0378-8733(89)90017-8 (cit. on p. 3).

[59] B. Jin, L. Liang, R. Rousseau, and L. Egghe. "The R-and AR-indices: Complementing the h-index." In: *Chinese science bulletin* 52.6 (2007), pp. 855–863 (cit. on pp. 81, 83, 84).

[60] R. Johnson, A. Watkinson, and M. Mabe. "The STM report." In: *International Association of Scientific, Technical, and Medical Publishers*. 2018 (cit. on p. 11).

[61] R. Johnson, A. Watkinson, and M. Mabe. *The STM report*. Tech. rep. International Association of Scientific, Technical, and Medical Publishers, 2018 (cit. on p. 37).

[62] I. Kecskés and L. R. Horn. *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects*. Vol. 1. Walter de Gruyter, 2008 (cit. on p. 38).

[63] H. Khayrallah, B. Thompson, K. Duh, and P. Koehn. "Regularized training objective for continued training for domain adaptation in neural machine translation." In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. 2018, pp. 36–44 (cit. on p. 56).

[64] S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin. "Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles." In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. 2010, pp. 21–26 (cit. on pp. 45–47).

[65] M. Kosmulski et al. "A new Hirsch-type index saves time and works equally well as the original h-index." In: *ISSI newsletter* 2.3 (2006), pp. 4–6 (cit. on p. 85).

[66] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289 (cit. on p. 15).

[67] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. “Albert: A lite bert for self-supervised learning of language representations.” In: *arXiv preprint arXiv:1909.11942* (2019) (cit. on pp. 65–69, 76).

[68] A. Lauscher, K. Eckert, L. Galke, A. Scherp, S. T. R. Rizvi, S. Ahmed, A. Dengel, P. Zumstein, and A. Klein. “Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph.” In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL ’18. Fort Worth, Texas, USA: ACM, 2018, pp. 109–118. DOI: 10.1145/3197026.3197050 (cit. on pp. 11, 16).

[69] V. I. Levenshtein. “Binary Codes Capable of Correcting Deletions, Insertions and Reversals.” In: *Soviet Physics Doklady* 10 (Feb. 1966), p. 707 (cit. on p. 44).

[70] Y. Li, T. Baldwin, and T. Cohn. “What’s in a domain? learning domain-robust text representations using adversarial training.” In: *arXiv preprint arXiv:1805.06088* (2018) (cit. on p. 56).

[71] C. Lin and Y. He. “Joint sentiment/topic model for sentiment analysis.” In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, pp. 375–384 (cit. on p. 53).

[72] *Linked Open Citation Database (LOC-DB)*. https://locdb.bib.uni-mannheim.de/blog/en/ (cit. on p. 16).

[73] M. Litvak, M. Last, H. Aizenman, I. Gobits, and A. Kandel. “DegExt—A language-independent graph-based keyphrase extractor.” In: *Advances in Intelligent Web Mastering–3*. Springer, 2011, pp. 121–130 (cit. on pp. 37, 38, 47).

[74] Z. Liu, P. Li, Y. Zheng, and M. Sun. “Clustering to find exemplar terms for keyphrase extraction.” In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. 2009, pp. 257–266 (cit. on pp. 45, 47).

[75] J. Long, E. Shelhamer, and T. Darrell. “Fully convolutional networks for semantic segmentation.” In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965 (cit. on pp. 16, 23).

[76] P. Lopez. “GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications.” In: *Research and Advanced Technology for Digital Libraries*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 473–474 (cit. on pp. 16, 29).

[77] P. Lopez and L. Romary. “HUMB: Automatic key term extraction from scientific articles in GROBID.” In: *Proceedings of the 5th international workshop on semantic evaluation*. Association for Computational Linguistics. 2010, pp. 248–251 (cit. on p. 47).

[78] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. "Learning Word Vectors for Sentiment Analysis." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. URL: http://www.aclweb.org/anthology/P11-1015 (cit. on p. 58).

[79] D. Mahata, R. R. Shah, J. Kuriakose, R. Zimmermann, and J. R. Talburt. "Theme-weighted Ranking of Keywords from Text Documents using Phrase Embeddings." In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2018, pp. 184–189. DOI: 10.31219/osf.io/tkvap (cit. on pp. 37, 45–47).

[80] D. Matsuoka, M. Ohta, A. Takasu, and J. Adachi. "Examination of effective features for CRF-based bibliography extraction from reference strings." In: *2016 Eleventh International Conference on Digital Information Management (ICDIM)*. Sept. 2016, pp. 243–248. DOI: 10.1109/ICDIM.2016.7829774 (cit. on p. 15).

[81] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. "Image-based recommendations on styles and substitutes." In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 2015, pp. 43–52 (cit. on p. 58).

[82] W. Medhat, A. Hassan, and H. Korashy. "Sentiment analysis algorithms and applications: A survey." In: *Ain Shams engineering journal* 5.4 (2014), pp. 1093–1113 (cit. on p. 53).

[83] D. Mercier, A. Bhardwaj, A. Dengel, and S. Ahmed. "SentiCite: An Approach for Publication Sentiment Analysis." In: *arXiv preprint arXiv:1910.03498* (2019) (cit. on pp. 56, 65, 76).

[84] J. L. Mey. *Whose language?: a study in linguistic pragmatics*. Vol. 3. John Benjamins Publishing, 1985 (cit. on p. 38).

[85] R. Mihalcea and P. Tarau. "Textrank: Bringing order into text." In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004 (cit. on pp. 37, 42, 47).

[86] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space." In: *CoRR* abs/1301.3781 (2013) (cit. on p. 43).

[87] P. Mongeon and A. Paul-Hus. "The journal coverage of Web of Science and Scopus: a comparative analysis." In: *Scientometrics* 106.1 (Jan. 2016), pp. 213–228. DOI: 10.1007/s11192-015-1765-5 (cit. on p. 11).

[88] N. Mrkšić, D. O. Séaghdha, B. Thomson, M. Gašić, P.-H. Su, D. Vandyke, T.-H. Wen, and S. Young. "Multi-domain dialog state tracking using recurrent neural networks." In: *arXiv preprint arXiv:1506.07190* (2015) (cit. on p. 57).

[89] M. Munikar, S. Shakya, and A. Shrestha. "Fine-grained Sentiment Classification using BERT." In: *2019 Artificial Intelligence for Transforming Business and Society (AITB)*. Vol. 1. 2019, pp. 1–5 (cit. on p. 55).

[90] G. Nikolentzos, P. Meladianos, Y. Stavrakas, and M. Vazirgiannis. "K-clique-graphs for Dense Subgraph Discovery." In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2017, pp. 617–633 (cit. on p. 37).

[91] B. O'Connor, M. Krieger, and D. Ahn. "Tweetmotif: Exploratory search and topic summarization for twitter." In: *AAAI*. 2010 (cit. on p. 42).

[92] Y. Ohsawa, N. E. Benson, and M. Yachida. "KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor." In: *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98-*. IEEE. 1998, pp. 12–18. DOI: 10.1109/adl.1998.670375 (cit. on p. 37).

[93] F. Osborne and E. Motta. "Klink-2: integrating multiple web sources to generate semantic topic networks." In: *ISWC*. Springer. 2015, pp. 408–424 (cit. on p. 43).

[94] F. Osborne, E. Motta, and P. Mulholland. "Exploring scholarly data with rexplore." In: *International semantic web conference*. Springer. 2013, pp. 460–477 (cit. on p. 43).

[95] L. Page, S. Brin, R. Motwani, and T. Winograd. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999 (cit. on p. 41).

[96] B. Pang and L. Lee. "A sentimental education: Sentiment analysis using subjectivity." In: *Proceedings of ACL*. 2004, pp. 271–278 (cit. on p. 58).

[97] T. Pay and S. Lucci. "Automatic Keyword Extraction: An Ensemble Method." In: *Conference: IEEE Big Data 2017, At Boston*. Dec. 2017 (cit. on p. 37).

[98] E. Kunnas. *PDFSSA4MET*. https://github.com/eliask/pdfssa4met (cit. on p. 14).

[99] A. Prasad, M. Kaur, and M.-Y. Kan. "Neural ParsCit: a deep learning-based reference string parser." In: *International Journal on Digital Libraries* 19.4 (2018), pp. 323–337 (cit. on p. 15).

[100] G. Rabby, S. Azad, M. Mahmud, K. Z. Zamli, and M. M. Rahman. “A Flexible Keyphrase Extraction Technique for Academic Literature.” In: *Procedia Computer Science* 135 (2018), pp. 553–563 (cit. on p. 37).

[101] S. Ren, K. He, R. Girshick, and J. Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” In: *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015, pp. 91–99 (cit. on p. 21).

[102] S. T. R. Rizvi, A. Lucieri, A. Dengel, and S. Ahmed. “Benchmarking Object Detection Networks for Image Based Reference Detection in Document Images.” In: *2019 Digital Image Computing: Techniques and Applications (DICTA)*. 2019, pp. 1–8 (cit. on p. 16).

[103] S. T. R. Rizvi, A. Dengel, and S. Ahmed. “A Hybrid Approach and Unified Framework for Bibliographic Reference Extraction.” In: *IEEE Access* 8 (2020). DOI: 10.1109/ACCESS.2020.3042455 (cit. on pp. 13, 17, 18, 21–25, 27–29).

[104] F. Rousseau and M. Vazirgiannis. “Main core retention on graph-of-words for single-document keyword extraction.” In: *European Conference on Information Retrieval*. Springer. 2015, pp. 382–393 (cit. on pp. 37, 47).

[105] R. Rousseau. “New developments related to the Hirsch index.” In: (2006) (cit. on p. 82).

[106] H. Sajjad, N. Durrani, F. Dalvi, Y. Belinkov, and S. Vogel. “Neural machine translation training in a multi-domain scenario.” In: *arXiv preprint arXiv:1708.08712* (2017) (cit. on p. 56).

[107] A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, and E. Motta. “The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II.” In: *The Semantic Web – ISWC 2018*. Springer, Jan. 2018, pp. 187–205 (cit. on pp. 38, 42).

[108] A. A. Salatino, F. Osborne, T. Thanapalasingam, and E. Motta. “The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles.” In: *Digital Libraries for Open Knowledge*. Cham: Springer International Publishing, 2019, pp. 296–311 (cit. on p. 43).

[109] G. Sautter and K. Böhm. “Improved Bibliographic Reference Parsing Based on Repeated Patterns.” In: *Theory and Practice of Digital Libraries*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 370–382 (cit. on pp. 11, 14).

[110] S. Scholar. *Semantic Scolar*. URL: https://www.semanticscholar.org/ (visited on 06/30/2022) (cit. on p. 4).

[111] M. Schreiber. “The influence of self-citation corrections and the fractionalised counting of multi-authored manuscripts on the Hirsch index.” In: *Annalen der Physik* 18.9 (2009), pp. 607–621 (cit. on p. 85).

[112] *Science Parse*. https://github.com/allenai/science-parse (cit. on p. 15).

[113] C. H. Sekercioglu. “Quantifying coauthor contributions.” In: *Science* 322.5900 (2008), pp. 371–371 (cit. on pp. 85, 86).

[114] K. Slabbekoorn, T. Noro, and T. Tokuda. “Ontology-Assisted Discovery of Hierarchical Topic Clusters on the Social Web.” In: *J. Web Eng.* 15.5&6 (2016), pp. 361–396 (cit. on p. 42).

[115] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. “Recursive deep models for semantic compositionality over a sentiment treebank.” In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, pp. 1631–1642 (cit. on p. 58).

[116] D. Su, Y. Xu, G. I. Winata, P. Xu, H. Kim, Z. Liu, and P. Fung. “Generalizing question answering system with pre-trained language model fine-tuning.” In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 2019, pp. 203–211 (cit. on p. 56).

[117] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. “Learning sentiment-specific word embedding for twitter sentiment classification.” In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 1555–1565 (cit. on p. 55).

[118] T. Thongtan and T. Phienthrakul. “Sentiment Classification Using Document Embeddings Trained with Cosine Similarity.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 407–414. DOI: 10.18653/v1/P19-2057 (cit. on p. 55).

[119] D. Tkaczyk, R. Gupta, R. Cinti, and J. Beel. “ParsRec: A Novel Meta-Learning Approach to Recommending Bibliographic Reference Parsers.” In: *ArXiv* abs/1811.10369 (2018) (cit. on p. 15).

[120] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski. “CERMINE: automatic extraction of structured metadata from scientific literature.” In: *International Journal on Document Analysis and Recognition (IJDAR)* 18.4 (Dec. 2015), pp. 317–335. DOI: 10.1007/s10032-015-0249-8 (cit. on p. 15).

[121] R. Wang, W. Liu, and C. McDonald. “Using word embeddings to enhance keyword identification for scientific publications.” In: *ADC*. Springer. 2015, pp. 257–268 (cit. on pp. 45–47).

[122] M. Ware and M. Mabe. "The STM report." In: *International Association of Scientific, Technical, and Medical Publishers*. 2015 (cit. on p. 11).

[123] M. Ware and M. Mabe. *The STM report: An overview of scientific and scholarly journal publishing*. Tech. rep. International Association of Scientific, Technical, and Medical Publishers, 2015 (cit. on p. 37).

[124] J. D. West, M. C. Jensen, R. J. Dandrea, G. J. Gordon, and C. T. Bergstrom. "Author-Level Eigenfactor Metrics: Evaluating the influence of authors, institutions and countries within the SSRN community." In: *Harvard Business School NOM Unit Working Paper* 12-068 (2012) (cit. on p. 87).

[125] Z. Wu, Y. Rao, X. Li, J. Li, H. Xie, and F. L. Wang. "Sentiment detection of short text via probabilistic topic modeling." In: *International Conference on Database Systems for Advanced Applications*. Springer. 2015, pp. 76–85 (cit. on p. 53).

[126] Q. Xie, Z. Dai, E. H. Hovy, M. Luong, and Q. V. Le. "Unsupervised Data Augmentation." In: *CoRR* abs/1904.12848 (2019). arXiv: 1904.12848 (cit. on p. 55).

[127] J. Xu, Y. Zhang, Y. Wu, J. Wang, X. Dong, and H. Xu. "Citation sentiment analysis in clinical trial papers." In: *AMIA annual symposium proceedings*. Vol. 2015. American Medical Informatics Association. 2015, p. 1334 (cit. on pp. 55, 59).

[128] E. Yan and Y. Ding. "Discovering author impact: A PageRank perspective." In: *Information processing & management* 47.1 (2011), pp. 125–134 (cit. on p. 87).

[129] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. "Xlnet: Generalized autoregressive pretraining for language understanding." In: *Advances in neural information processing systems*. 2019, pp. 5754–5764 (cit. on pp. 63–65, 76).

[130] C.-T. Zhang. "A proposal for calculating weighted citations based on author rank." In: *EMBO reports* 10.5 (2009), pp. 416–417 (cit. on p. 86).

[131] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu. "Attention-based bidirectional long short-term memory networks for relation classification." In: *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*. 2016, pp. 207–212 (cit. on p. 55).

[132] *Zotero*. https://www.zotero.org/ (cit. on p. 30).

[133] J. Zou, D. Le, and G. R. Thoma. "Locating and parsing bibliographic references in HTML medical articles." In: *International Journal on Document Analysis and Recognition (IJDAR)* 13.2 (June 2010), pp. 107–119. DOI: 10.1007/s10032-009-0105-9 (cit. on p. 15).

INDEX

ACADEMIC CURRICULUM VITÆ: **SYED TAHSEEN RAZA RIZVI**

CONTACT INFORMATION

Email: syed_tahseen_raza.rizvi@dfki.de

EDUCATION

University of Kaiserslautern-Landau, Germany — *2017 - 2023*
PhD Student
PhD Topic: *From Scientific Publications to Community Insights*

Technische Universität Kaiserslautern, Germany — *2014 - 2016*
Master of Science in Computer Science
Master Thesis: *Information Extraction from Technical Documents*

Punjab University College of Information Technology, Lahore — *2008 - 2012*
Bachelor of Science in Software Engineering

EXPERIENCE

German Research Center for Artificial Intelligence (DFKI) — *since 03/2016*
Smart Data and Knowledge Services Department, Prof. Dengel
Main Projects: LOCDB, KI-Wissen

Kyushu University, Japan — *2016 - 2017*
Visiting Researcher

Bosch, Germany — *2015 - 2016*
Internship & MS Thesis

Technische Universität Kaiserslautern — *2015 - 2016*
Scientific Assistant

eGenienext Web Solutions, Pakistan — *2013 - 2014*
Software Engineer

PUBLICATIONS

Utilizing Out-Domain Datasets to Enhance Multi-Task Citation Analysis — *Springer AAI 2022*
D. Mercier, **STR. Rizvi**, V. Rajashekar, S. Ahmed, A. Dengel

ImpactCite: An XLNet-based method for Citation Impact Analysis — *ICAART 2021*
D. Mercier, **STR. Rizvi**, V. Rajashekar, A. Dengel, S. Ahmed

IEEE Access 2020 A Hybrid Approach and Unified Framework for Bibliographic Reference Extraction
STR. Rizvi, A. Dengel, S. Ahmed

DAS 2020 From automatic keyword detection to ontology-based topic modeling
M Beck, **STR. Rizvi**, A. Dengel, S. Ahmed

Springer Nature CS 2020 Benchmarking deep learning models for classification of book covers
A. Lucieri, H. Sabir, SA. Siddiqui, **STR. Rizvi**, A. Dengel, S. Ahmed

DAS 2020 From automatic keyword detection to ontology-based topic modeling
M Beck, **STR. Rizvi**, A. Dengel, S. Ahmed

DICTA 2019 Benchmarking object detection networks for image based reference detection in document images
STR. Rizvi, A. Lucieri, A. Dengel, S. Ahmed

DICTA 2019 FFD: Figure and formula detection from document images
J. Younas, **STR. Rizvi**, MI. Malik, F. Shafait, P. Lukowicz, S. Ahmed

ICDAR 2019 DeepTabStR: deep learning based table structure recognition
SA. Siddiqui, IA. Fateh, **STR. Rizvi**, A. Dengel, S. Ahmed

JCDL 2018 Linked open citation database: Enabling libraries to contribute to an open and interconnected citation graph
A. Lauscher, K. Eckert, L. Galke, A. Scherp, STR. Rizvi, S. Ahmed, A. Dengel, P. Zumstein, A. Klein

ICAART 2018 Ontology-based Information Extraction from Technical Documents
STR. Rizvi, D. Mercier, S. Agne, S. Erkel, A. Dengel, S. Ahmed