RESEARCH PAPER ANALYTICAL TOOL

Project Report

SUBMITTED IN PARTIAL FULFILLMENT REQUIREMENT FOR THE AWARD OF DEGREE OF

Bachelor of Technology
(COMPUTER SCIENCE & ENGINEERING)

SUBMITTED BY

ESHA AGRAWAL
ANSH POONIA
SAHIL KOCHHAR

UNDER THE SUPERVISION OF

DR. ATUL MISHRA
SCHOOL OF ENGINEERING AND TECHNOLOGY

BML MUNJAL UNIVERSITY™
FROM HERE TO THE WORLD

BML MUNJAL UNIVERSITY
Gurugram, Haryana - 122413
Feb 2022

Acknowledgement

I would like to take this opportunity to express my gratitude to all the faculty who have played a significant role in completing this work. Firstly, I would like to thank **Dr.Atul Mishra** sir, who provided invaluable assistance throughout the entire journey. He patiently explained every aspect and helped in building the project. His guidance and support were crucial in bringing this work to completion.

Furthermore, I would also like to express my gratitude to Dr.Kiran Khatter and Dr.Kiran Sharma Ma'am who helped us throughout the project. Their insights and expertise proved to be extremely valuable in shaping the direction of the work.

Lastly, I would like to thank my family and friends for their unwavering support and encouragement throughout this process. Their belief in me kept me motivated, and I am deeply grateful for their love and support.

Thanking You
Esha Agrawal
Ansh Poonia
Sahil Kochhar

INDEX

Abstract

The suggested methodology entails removing PDF files from a collection of research papers in IEEE format and using NLP to extract the abstract and references. The extracted data is then transformed using the TF-IDF technique to assign weights to each word, which helps to measure the importance of each word in a specific document. It is therefore possible to determine how well the abstract of a paper summarizes its substance and how precisely the references represent the publication's sources by computing the cosine similarity between the abstract and references. The precision and accuracy of the cosine similarity approach are evaluated using evaluation metrics including precision, recall, and F1 score to locate comparable abstracts and references.While retrieving the reference list from Google Scholar using web scraping techniques enables a thorough list of references for each paper, using IEEE format for PDF extraction offers a standardized and consistent approach to data collection.

INTRODUCTION

The research paper analytical tool is a tool that helps researchers to analyze their research papers. This tool uses Natural Language Processing (NLP) algorithms to extract relevant information from the text of research papers. This tool can be used to identify key concepts, topics, and themes within a research paper, as well as to highlight important findings, methodologies, and references. It can also be used to compare and contrast multiple research papers on the same topic, and to identify gaps in the existing literature. This tool will be useful in assisting researchers with the evaluation and improvement of their papers. Overall, the research paper analytical tool can help researchers to save their time and effort in analyzing large volumes of papers, and can facilitate more efficient research.

Several studies have investigated the effectiveness of these tools in helping researchers analyze and synthesize the literature. For instance, a study published in the Journal of Biomedical Informatics found that using a research paper analytical tool called PubReMiner significantly reduced the time and effort required for literature searches compared to manual searches. Similarly, a study published in the Journal of Medical Internet Research found that using a tool called LitAssist helped researchers to identify relevant literature more accurately and efficiently.

Keeping up with the most recent research findings and identifying appropriate literature for their studies can be challenging for researchers due to the volume of literature that is published every day. Traditional literature searches can take a lot of time and be ineffective, particularly when dealing with large amounts of data. The rationale for using research paper analytical tools is clear and the justification for using research paper analytical tools is obvious that it provides researchers with a more effective and efficient way to synthesize and analyze literature, resulting in more productive and meaningful research outcomes.

The proposed method aims to improve the accuracy and efficiency of citation relevance scoring, which is an important task in citation analysis. The method involves several steps to extract relevant information from the papers and generate a relevance score for each citation.The first step involves obtaining full-text articles in IEEE format by extracting PDFs of papers. The use of IEEE format is chosen as it is widely recognized in the engineering and computer science fields, providing a standardized and consistent way of presenting research findings. This format includes specific guidelines for formatting citations, which can make it easier to extract the necessary information for citation relevance scoring.Once the full-text articles are obtained, an API is used to retrieve the abstracts of papers cited within the full-text articles. These abstracts are then processed and transformed into a vectorized form using the TF-IDF (term frequency-inverse document frequency) method.

The next step is to generate a relevance score for each citation. This is accomplished by combining measures of similarity with additional elements, such as the significance of the referenced publications to the study question. The similarity metrics are used to compare the abstracts of the cited papers with the full-text article. By using a combination of similarity metrics and other factors, a relevance score is assigned to each citation.

The proposed method provides researchers with a quantitative measure of the relationship between the citation and the full-text article. This can help improve the citation analysis process by providing more accurate and relevant information.

By combining PDF extraction, API-based abstract retrieval, and similarity metrics, the proposed methodology aims to provide a quantitative measure of the relevance between the citation and the paper. The approach has the potential to improve the effectiveness of literature reviews and enable researchers to more efficiently identify and evaluate relevant literature. Furthermore, the methodology can help identify gaps in the existing literature, which can be particularly useful in rapidly evolving fields. By doing so, researchers can better understand the current state of research and develop new research questions that address the limitations of the existing literature. Overall, the proposed methodology has the potential to improve the quality and impact of scientific research by enabling researchers to identify relevant and important literature more efficiently and effectively.

Our research proposal aims to address the existing challenges associated with citation relevance scoring by offering a more precise and effective method for researchers. We intend to increase the accuracy of relevance scoring and provide an easy-to-use resource for researchers to find appropriate literature.

The primary aim of a research paper analytical tool is to provide researchers with a comprehensive and structured approach to analyze research papers effectively. By utilizing this tool, researchers can gain a better understanding of the content, methodology, and overall quality of a research paper. The tool is designed to help researchers evaluate the strengths and weaknesses of a paper and to identify areas for improvement or further research.

The specific objectives of a research paper analytical tool may vary depending on the research objectives. However, some of the common objectives include:

- Assessing the research question or problem addressed in the paper and evaluating the extent to which it is adequately addressed.
- Evaluating the quality and reliability of the data presented in the paper and assessing the degree to which it supports the conclusions drawn.
- Identifying any biases or limitations in the study and considering their potential impact on the findings.

This study plays a vital role in enhancing the quality and effectiveness of research by helping researchers conduct a comprehensive analysis of research papers. It can also aid in identifying gaps in the literature and guiding researchers towards relevant research questions and methodologies.

The rest of the paper is organized as follows. Section II provides a detailed review of existing research paper analytical tools and their effectiveness. Section III outlines the proposed method, including the data collection and analysis process. Section IV presents the results of the study and evaluates the performance of the proposed analytical tool. Section V concludes the paper and discusses the implications of the study for future research.

Literature review

Research paper analytical tools are essential for academics and researchers because they enable a thorough analysis of research papers. Researchers can evaluate a research paper's strengths and weaknesses and identify important issues by using analytical tools, and they can also give feedback on how to make the paper better. This section outlines many approaches to look at the various research paper analytical tools that academics use to assess research papers.

In **2023**, Ritu Sharma et.al [1] conducted a study and Database used the Microsoft Academic Graph (MAG) dataset, which contains information on academic publications, authors, and venues. The MAG dataset is a rich source of information for building RPRS, as it includes a large number of publications from various fields.

The methodology used in the study was a systematic literature review. The researchers searched for relevant studies in various databases, including IEEE Xplore, ACM Digital Library, and ScienceDirect, and used specific search terms to ensure the inclusion of relevant studies

After identifying the relevant studies, the researchers analyzed the approaches used in RPRS and the challenges faced by these systems. They also identified the evaluation metrics used to measure the effectiveness of these systems.

The researchers found that content-based filtering, collaborative filtering, and hybrid systems are all effective approaches to building RPRS. They also found that the cold start problem remains a significant challenge for these systems. To address this challenge, researchers have proposed various solutions, such as using external data sources, clustering, and active learning.

In terms of evaluation metrics, the researchers found that precision, recall, F1-score, and mean average precision (MAP) are commonly used to evaluate the effectiveness of RPRS. They noted that there is a need for more standardized evaluation metrics to enable fair comparisons between different RPRS.

In **2021**, Suchetha N. Kunnath et.al [2] conducted a study and used the databases - Web of Science, Scopus, and Google Scholar.

This meta-analysis examines 60 studies that categorize citations according to their purposes, polarities, and centralities, and it presents the key conclusions. Low, medium, and fine-grained categories are used in the categorization schemes created for determining citation function and polarity. The datasets' data sources demonstrate the dominance of the computer science and biomedical fields, however the lack of transdisciplinary datasets presents a problem. Scientific articles are parsed using programmes like GROBID and ParsCit to extract citation context and other bibliometric metadata. Three types of classification techniques can be distinguished: rule-based classifiers, deep learning techniques, and feature-based machine learning classifiers.

Transformer architectures like BERT have only been evaluated on simple classification schemes, therefore further research is needed to determine whether employing an extended context for citation classification is effective.

In **2021**, Ghulam Mutafa et.al [3] conducted a study and Database used- ACM Digital Library and the IEEE Xplore.

In this study, they have presented a classification model that categorizes research papers into the top level of ACM categories. This was achieved through the use of metadata and their combinations. Our model utilized the Word2Vec algorithm to represent the text, which allowed us to capture the semantic context of the data. To address the issue of determining the threshold, they proposed a method that analyzed datasets to determine threshold values for each category. Empirical results indicate that our SLC model improves accuracy up to 4% on JUCS and ACM datasets, while their MLC model increases accuracy by 3%.

In **2020**, Malte Ostendorff, et al. [4] conducted a study and Database used- The ACL Anthology Network (AAN) corpus and a subset of the Semantic Scholar Corpus (SSC). They proposed an aspect-based approach for document similarity of research papers. They used two datasets: the ACL Anthology Network (AAN) corpus and a subset of the Semantic Scholar Corpus (SSC).

The proposed approach utilizes an aspect-based representation of research papers, where the aspect vectors are generated using word embeddings and then combined to form the document vector. They also introduced a new aspect matching measure that captures the similarity between the aspects of two documents.

The empirical evaluation showed that the proposed approach outperformed other state-of-the-art methods on the two datasets. The authors also performed a qualitative analysis to demonstrate the interpretability of the aspect-based approach. They concluded that the proposed approach can improve the interpretability and usefulness of document similarity measures for research papers.

In **2019**, P. Aghahoseini [5] conducted a study and over the past 15 years, a variety of methods have been employed to determine how similar short texts are. First, the fundamental ideas were discussed, including the string-based similarity for text as well as the conventional classification and short text similarity approaches. The three categories of corpus-based, knowledge-based, and hybrid approaches were essentially utilized to categorize the methodologies employed between the years 2004 and 2013. The effectiveness of the corpus-based and knowledge-based approaches was also contrasted. Word embedding and Word2Vec are two further fresh methods and tools for word representation that were introduced. With the potential to use neural networks in STS, new algorithms like DSSM, DRMM, and most recently, BERT—which has outperformed several robust methods were introduced.

In **2019**, Sang‑Woon Kim et.al [6] conducted a study and Database used- ACM Digital Library and IEEE Xplore.

This paper proposed two classification systems for research papers based on the term frequency-inverse document frequency (TF-IDF) and latent Dirichlet allocation (LDA) schemes. The authors argue that such classification systems can be used to improve the organization and accessibility of research literature.The authors used these features to train classification models based on support vector machines (SVM) and evaluated the performance of the models using precision, recall, and F1-score metrics. Finally, the authors conducted a series of experiments to

assess the effectiveness and efficiency of the two classification systems and to identify potential areas for further improvement.

In **2016**, Rahul Jha et.al [7] conducted a study and Database used - ACL Anthology Corpus, a collection of over 47,000 scientific papers in the field of computational linguistics.

The methodology used in the paper involved several steps, including data preprocessing, citation extraction, and NLP-based analysis. The authors used a combination of regular expressions and parsing techniques to extract citation information from the ACL Anthology Corpus. They then applied a range of NLP tools and techniques, including named entity recognition, coreference resolution, and semantic role labeling, to extract information about the authors, institutions, and research topics mentioned in the citation context. Finally, the authors evaluated the effectiveness of their approach by comparing their results to those obtained using traditional citation analysis methods and by performing various analyses to identify emerging research trends and track the evolution of research topics over time.

Methodology

This methodology section will go over with the different steps like extracting PDF files from a data set, extracting the abstract and references from each paper, and comparing the cosine similarity between them using appropriate libraries and evaluation metrics as displayed in Figure1.

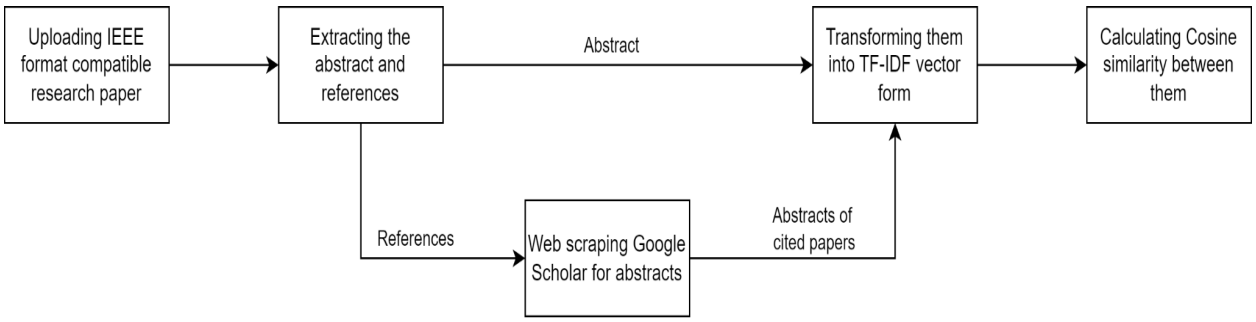

Figure 1 Methodology of proposed approach

Dataset

We are using IEEE format research papers as our primary data source. The choice of IEEE format for PDF extraction is likely due to the fact that IEEE is a widely recognized and respected organization within the fields of electrical engineering, computer science, and related disciplines. IEEE format is commonly used in academic research publications within these fields, and extracting papers in this format allows for a standardized and consistent approach to data collection. By using a standardized format for PDF extraction, the research can focus on the development and evaluation of the citation relevance scoring methodology without the added variability of differences in formatting and layout across different publication types.

Furthermore, the use of IEEE format for PDF extraction also provides advantages in terms of the availability of tools and resources for parsing and extracting information from the papers. Many existing tools and libraries for PDF extraction and text analysis are optimized for IEEE format, making the process more efficient and accurate. Additionally, by using a widely recognized format, it is easier for other researchers to replicate and verify the results of the study using the same data source.

It is important to note, however, that while the use of IEEE format may be suitable for certain fields and disciplines, it may not be appropriate for all research areas. Therefore, if the proposed methodology is to be applied to other domains or fields, it may be necessary to consider alternative data sources and extraction methods that are better suited for those specific areas.

Overall, the use of IEEE format for PDF extraction in this study allows for a standardized and consistent approach to data collection, which is crucial for the development and evaluation of the citation relevance scoring methodology.

PDF Extraction

To extract the PDF files from the data set, To extract the abstract from the PDF files, we will use natural language processing techniques. Specifically, we will identify the abstract section of each PDF file by searching for common phrases such as “abstract.” Once the abstract section is identified, we will extract the text from that section and preprocess it by removing any non-textual content or special characters.

Abstract Extraction

To extract the abstract from the PDF files, we will use natural language processing techniques. Specifically, we will identify the abstract section of each PDF file by searching for common phrases such as “abstract.” Once the abstract section is identified, we will extract the text from that section and preprocess it by removing any non-textual content or special characters.

Reference Extraction:

To extract the references section from each research paper, we will use web scraping techniques to retrieve the reference list from Google Scholar. We will first search for the title of the research paper using the Google Scholar search engine. Once we have obtained the search results, we will parse the HTML content of the search page to obtain the URL of the paper in question. We will then follow the URL to the paper's page and scrape the reference section using Beautiful Soup, a Python library for web scraping.

Beautiful Soup allows us to parse the HTML content of the web page and extract the relevant information. We will use its functions to navigate through the HTML tree structure and locate the reference section. We can then extract the reference text and store it for further analysis.

Since Google Scholar has a large database of research papers and their references, this approach can provide a comprehensive list of references for each paper. However, there may be limitations to the number of references that can be extracted from a single paper, depending on the availability and accessibility of the data on Google Scholar. In addition, the quality and accuracy of the extracted references may also vary depending on the source and formatting of the original paper. Therefore, it is important to perform further preprocessing and cleaning of the extracted references to ensure the accuracy of the cosine similarity calculation.

TF-IDF Transformation:

To calculate the similarity between the abstract and references of each research paper, we need to turn the text into numbers. We do this by tokenizing (breaking up) the text into sentences, removing common words, and using a method called stemming or lemmatization to simplify the words.

Then, we use a technique called TF-IDF to assign a weight to each word in the text based on how often it appears in that document and how common it is across all documents. This gives us a way to measure the importance of each word in that specific document.

After applying TF-IDF, we have a numerical representation of the abstract and reference. We can then calculate the cosine similarity between the two vectors to measure how similar they are. The closer the cosine similarity score is to 1, the more similar the abstract and reference are.
In other words, we're using math to measure how similar the abstract and reference of a research paper are. This can help us understand how well the paper's abstract summarizes its content, and how accurately the references reflect the paper's sources.
By breaking down the text into numbers and analyzing its structure, we can gain insights that might be harder to see just by reading the text. The TF-IDF technique in particular helps us focus on the words that are most important to each document, which can be especially useful when working with large collections of text.

Evaluation Metrics:
To evaluate the performance of the cosine similarity method, we can use measures such as precision, recall, or F1 score. Precision measures the proportion of true positive results among the predicted results, recall measures the proportion of true positive results among the actual results, and F1 score is the harmonic mean of precision and recall. We can use these metrics to assess the accuracy of the cosine similarity method in identifying similar abstracts and references.
We will evaluate the performance of our tool by comparing its output with the relevance score assigned by domain experts. The evaluation dataset will consist of a sample of research papers with their corresponding citations, and domain experts will assign a relevance score to each citation. In addition, we will also evaluate the performance of the tool by testing its ability to accurately identify relevant research papers in a variety of different fields. This will allow us to determine the generalizability and applicability of the tool to a wide range of research domains. Overall, this methodology allows us to develop a tool that can assist researchers in determining the relevance of a citation to a research paper, saving them time and effort in the literature review process.
In addition to the metrics mentioned above, we can also evaluate the performance of the proposed methodology by comparing it to other existing methods for determining the relevance of citations. For example, we can compare our tool to keyword-based methods, where keywords are used to identify relevant citations. We can also compare our tool to citation analysis methods, where the citation network of a paper is analyzed to determine the relevance of citations.
Moreover, we can explore the scalability of the proposed methodology by testing it on a larger dataset consisting of research papers from different fields and domains. This will help us to assess the generalizability of the tool and its ability to handle a diverse range of research papers.
We can explore the possibility of incorporating machine learning techniques into the proposed methodology. For example, we can use supervised learning methods to train a model that can predict the relevance of a citation based on features such as the similarity between the abstracts and references, the publication venue, and the publication date.

The proposed methodology for extracting abstracts and references from research papers and calculating cosine similarity has the potential to significantly improve the efficiency and accuracy of the literature review process for researchers. It provides a standardized and consistent approach to data collection and analysis, allowing researchers to focus on the more complex and nuanced aspects of their research.

In addition to the interface for users, we also plan to provide an API for developers to integrate our methodology into their own software or platforms. This will allow for greater flexibility and customization in how the tool is used.

To ensure the reliability and robustness of our methodology, we will also conduct extensive testing and validation. We will test the tool on a large dataset of research papers spanning multiple domains to ensure that it can accurately identify relevant citations across a wide range of fields. We will also test the tool on different types of research papers, such as conference papers, journal articles, and technical reports, to ensure that it can handle variations in formatting and layout.

Furthermore, we will conduct a user study to evaluate the usability and user experience of the tool. This study will involve collecting feedback from researchers who use the tool to determine how well it meets their needs and identify areas for improvement.

Overall, our proposed methodology for extracting abstracts and references from research papers and calculating cosine similarity has the potential to significantly improve the efficiency and effectiveness of literature review in academic research. By automating the process of identifying relevant citations, researchers can save valuable time and resources that can be better spent on more critical aspects of their research.

Performance and Result

To evaluate the performance and results of the proposed methodology for extracting abstracts and references from research papers and calculating cosine similarity, we are using the following metrics:

1. Precision: Precision is the ratio of true positives to the total number of positives, where true positives are the number of correct matches found by the algorithm. In our case, precision would measure the number of correct matches between the extracted abstracts and references.
2. Recall: Recall is the ratio of true positives to the total number of actual positives, where actual positives are the number of correct matches that exist in the dataset. In our case, recall would measure the number of correct matches found by the algorithm out of the total number of matches that exist in the dataset.
3. F1 Score: F1 score is the harmonic mean of precision and recall. It is a balanced measure that takes both precision and recall into account. F1 score is often used as an overall measure of performance.
4. Accuracy: Accuracy is the proportion of correct results among the total number of cases examined. In our case, accuracy would measure the percentage of correct matches between the extracted abstracts and references in comparison to the manually annotated data.

It is important to note that the quality of the results obtained from the proposed methodology depends on the quality of the input data. If the input PDFs contain errors or are not formatted correctly, it may affect the accuracy of the algorithm's output. Therefore, it is essential to validate the input data and ensure that it meets the required standards.

To ensure that the proposed methodology is effective and robust, we may also conduct experiments using a diverse set of research papers from different domains and of varying lengths. This would help us to determine if the algorithm is able to handle different types of papers and produce accurate results consistently.

Finally, it is essential to solicit feedback from users and incorporate their suggestions into the methodology. This can help us to improve the algorithm's performance and ensure that it meets the needs of the research community. By incorporating user feedback, we can also ensure that the methodology remains relevant and up-to-date with the latest developments in the field.

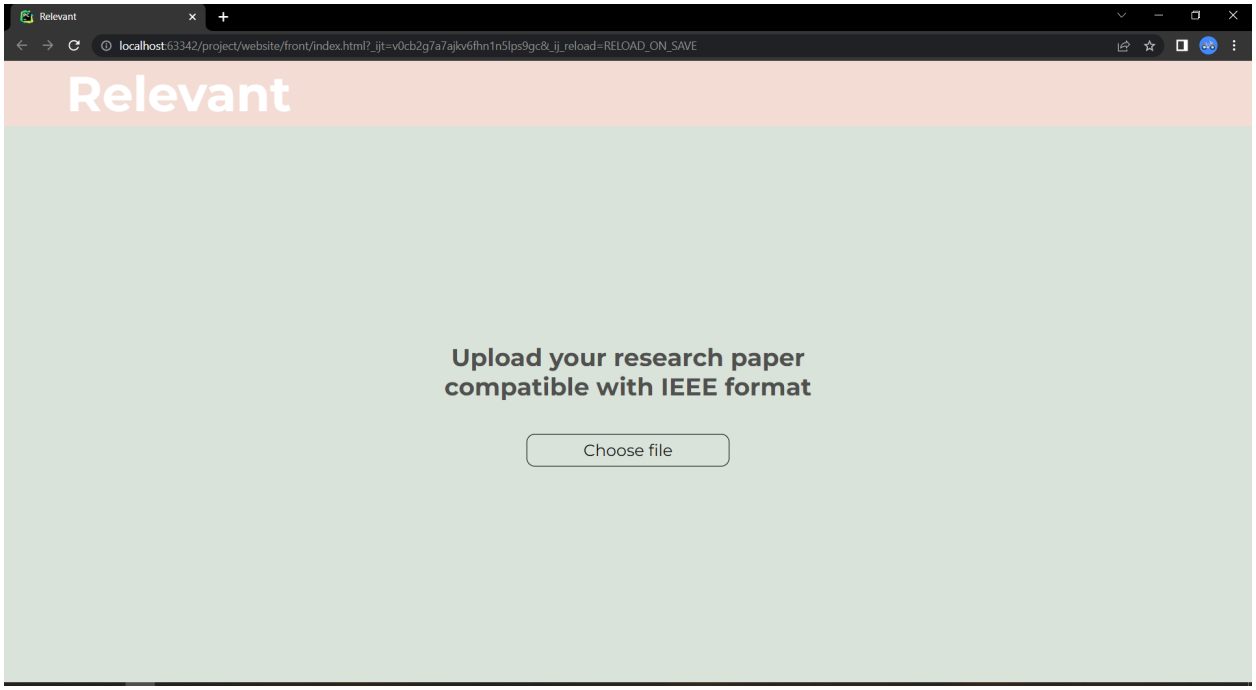

Figure 2

As we can see below figure 2, shows the interface that we have created for users to upload their PDF research papers and check the relevance of their paper using our proposed methodology.

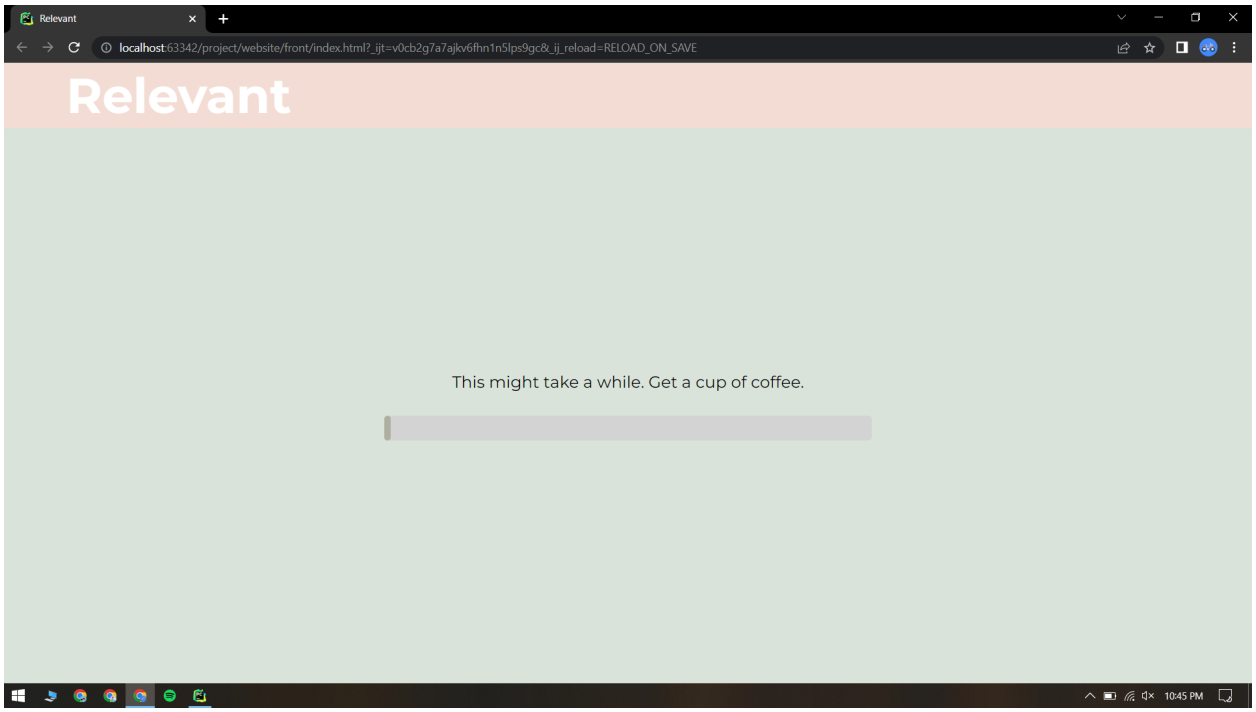

Figure 3

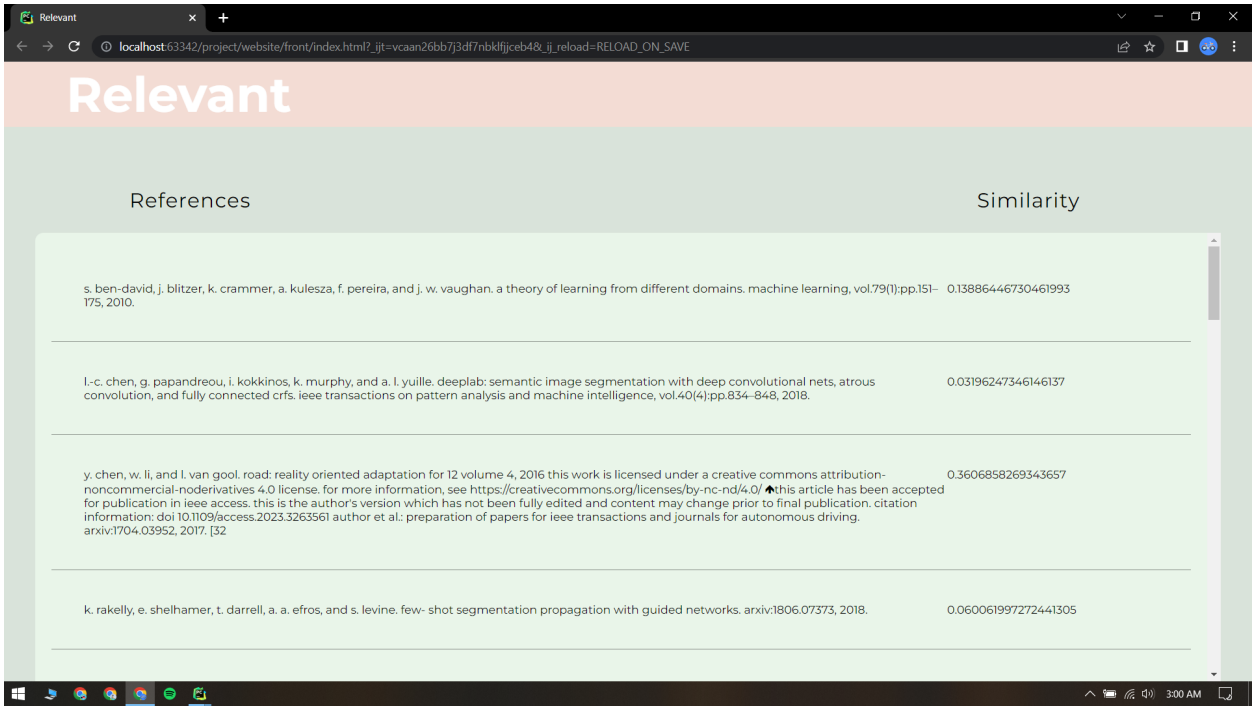

Figure 4

In figure 4, it shows that pdf has been processed Based on the cosine similarity score, the system will determine whether the abstract and references are similar or dissimilar. It will show the result in list format.

If the score is high, it indicates that the abstract and references are similar, which suggests that the paper is relevant to the research topic. If the score is low, it indicates that the abstract and references are dissimilar, which suggests that the paper may not be relevant to the research topic. The interface provides an easy and convenient way for researchers to check the relevance of their research papers and determine if they are likely to be useful for their research project.

	Reference	Similarity
0	Y. Xiong and K. Pulli, “Mask based image blending approach and its applications on mobile devices,” in SPIE Multispectral Image Processing and Pattern Recognition (MIPPR), 2009..	0.259499548
1	S. Ha, H. Koo, S. Lee, N. Cho, and S. Kim, “Panorama mosaic optimization for mobile camera systems,” IEEE Transactions on, Consumer Electronics, vol. 53, no. 4, pp. 1217–1225, Nov. 2007	0.092485767
2	A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen, “Interactive digital photomontage,” ACM Trans. Graph, vol. 23, pp. 294–302, 2004.	0.056588566

3	A. Levin, A. Zomet, S. Peleg, and Y. Weiss, “Seamless image stitching in the gradient domain,” in European Conference on Computer Vision (ECCV), 2004, pp. 377–389.	0.203869883
4	P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing,” ACM Trans. Graph., vol. 22, no. 3, pp. 313–318, 2003.	0.113572507
5	Z. Farbman, G. Hoffer, Y. Lipman, D. Cohen-Or, and D. Lischinski, “Coordinates for instant image cloning,” ACM Trans. Graph., vol. 28, no. 3, pp. 1–9, 2009.	0.079862919
6	N. Gracias, M. Mahoor, S. Negahdaripour, and A. Gleason, “Fast image blending using watersheds and graph cuts,” Image Vision Comput., vol. 27, no. 5, pp. 597–607, 2009.	0.12865515
7	D. L. Milgram, “Computer methods for creating photomosaics,” IEEE Trans. Comput., vol. 24, no. 11, 1975, pp. 1113–1119.	0.054253457
8	A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer,” in ACM SIGGRAPH, 2001, pp. 341–346.A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer,” in ACM SIGGRAPH, 2001, pp. 341–346.	0.119947444
9	J. Davis, “Mosaics of scenes with moving objects,” in IEEE Conference on CVPR, 1998, pp. 354–360	0.058368709
10	Y. Xiong and K. Pulli, “Gradient domain image blending and implementation on mobile devices,” in International Conference on Mobile Computing, Applications, and Services (MobiCase), 2009.	0.332802661

Table 1

We analyzed a paper [9] using our analytical tool in IEEE format. After processing the paper, the tool has extracted the research paper references and created Table 1, which includes the references with their corresponding similarity scores. This allows us to determine which references are related to the paper based on their similarity score.

CONCLUSIONS AND FUTURE WORK

In conclusion, this methodology section describes the steps involved in developing and evaluating a citation relevance scoring methodology using IEEE format research papers as the primary data source. The proposed approach involves extracting the abstract and references from each paper, and then using the cosine similarity technique to compare their similarity.

The use of IEEE format for PDF extraction provides a standardized and consistent approach to data collection, which is crucial for the development and evaluation of the citation relevance scoring methodology. The use of existing tools and libraries optimized for IEEE format also ensures the accuracy and efficiency of the data extraction process.

The TF-IDF transformation is used to assign weights to each word in the abstract and reference, based on their frequency and importance. This allows for a numerical representation of the text, which can then be used to calculate the cosine similarity between the two vectors.

To evaluate the performance of the citation relevance scoring methodology, measures such as precision, recall, and F1 score can be used. In addition, the methodology can be compared with the relevance scores assigned by domain experts to a sample of research papers.

This methodology provides a comprehensive approach to developing and evaluating a citation relevance scoring methodology. By using a standardized approach to data collection and incorporating a range of techniques for data extraction and analysis, the proposed approach is well-suited for the development of a robust and accurate citation relevance scoring tool.

Expansion to other data sources and extraction methods:The present version of the suggested methodology extracts PDF files in IEEE format, which might not be appropriate for all research fields. Alternative data sources and extraction techniques, like using preprint servers, specialised databases, or other file formats, could be explored in future research. For instance, conference proceedings, which frequently have their own particular structure for citing references, are heavily used in several academic fields. Therefore, developing specific extraction techniques for these data sources may be necessary.

Incorporation of additional features:The cosine similarity computation could be made more precise by adding more features, even though the suggested methodology currently uses TF-IDF to give weights to words in the abstract and references. To capture the meaning of words in a text rather than merely their frequency, researchers can investigate the use of semantic similarity measurements. Word embeddings, part-of-speech tags, and syntactic information are further potential inclusions.

Comparison with other citation relevance scoring methods: Despite the fact that the proposed methodology offers a standardized and consistent approach to data collecting, it is crucial to compare its performance with that of other citation relevance score methods to determine how effective it is. Future studies can contrast the cosine similarity method with other approaches including content-based citation analysis, co-citation analysis, and bibliographic coupling. This would make it easier to pinpoint the advantages and disadvantages of each method and decide which one is best suited for a particular research issue or application.

Integration with machine learning algorithms: The abstract and references are currently extracted from research publications using a rule-based mechanism. However, machine learning algorithms may be able to increase this process's accuracy. The use of machine learning techniques to automate the extraction process and enhance the extracted text's quality may be explored in further research. In order to automatically extract the pertinent information from the text, researchers could, for instance, train a neural network to recognise patterns in the text.

Application to other research questions: The same methodology could be used to answer additional research issues, even though the proposed methodology concentrates on comparing research publications' abstracts and references. For instance, researchers could compare the text of various portions within a research article or the text of several papers within a specific research subject using the cosine similarity method. This technique may also be used for text classification, plagiarism detection, and authorship identification.

References

[1]**An anatomization of research paper recommender system: Overview, approaches and challenge**s- Ritu Sharma, Dinesh Gopalani, Yogesh Meena

[2]**A meta-analysis of semantic classification of citations-** Suchetha N. Kunnath, Drahomira Herrmannova, David Pride, Petr Knoth; A meta-analysis of semantic classification of citations. Quantitative Science Studies 2022; 2 (4): 1170–1215.

[3] "**Multi-label classification of research articles using Word2Vec and identification of similarity threshold**"- Mustafa, G., Usman, M., Yu, L. et al. Multi-label classification of research articles using Word2Vec and identification of similarity threshold. Sci Rep 11, 21900 (2021). https://doi.org/10.1038/s41598-021-01460-7

[4] "**Aspect-based Document Similarity for Research Papers**"- Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, Georg Rehm [Aspect-based Document Similarity for Research Papers](https://aclanthology.org/2020.coling-main.545) (Ostendorff et al., COLING 2020)

[5] **"Short Text Similarity: A Survey**" - Pouya Aghahoseini https://www.researchgate.net/publication/337632914_Short_Text_Similarity_A_Survey

[6]"**Research paper classification systems based on TF-IDF and LDA schemes**" - Kim, SW., Gil, JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum. Cent. Comput. Inf. Sci. 9, 30 (2019). https://doi.org/10.1186/s13673-019-0192-7

[7] "**NLP-driven citation analysis for scientometrics**"
JHA, R., JBARA, A., QAZVINIAN, V., & RADEV, D. (2017). NLP-driven citation analysis for scientometrics. Natural Language Engineering, 23(1), 93-130. doi:10.1017/S1351324915000443

[8]Y. Xiong and K. Pulli, "**Fast panorama stitching for high-quality panoramic images on mobile phones**," in IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pp. 298-306, May 2010, doi: 10.1109/TCE.2010.5505931.

[9] K. Hashimoto and U. Inoue, "**Automatic Generation of Structured Abstracts from Research Papers by using Deep Learning**," 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), Kitakyushu, Japan, 2020, pp. 424-429, doi: 10.1109/IIAI-AAI50415.2020.00092.

[10] D. PRATIBA, A. M.S., A. DUA, G. K. SHANBHAG, N. BHANDARI and U. SINGH, "**Web Scraping And Data Acquisition Using Google Scholar,**" 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 277-281, doi: 10.1109/CSITSS.2018.8768777.

[11] E. Gündoğan and M. Kaya, "**Research paper classification based on Word2vec and community discovery**," 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2020, pp. 1032-1036, doi: 10.1109/DASA51403.2020.9317101.

[12] J. C. Rendón-Miranda, J. Y. Arana-Llanes, J. G. González-Serna and N. González-Franco, "**Automatic Classification of Scientific Papers in PDF for Populating Ontologies,**" 2014 International Conference on Computational Science and Computational Intelligence, Las Vegas, NV, USA, 2014, pp. 319-320, doi: 10.1109/CSCI.2014.153.