Mining Future Scopes from Research Manuscripts

Arun Bharathapu[1], Sudhakar Anakala[2], Varun Paul Gade[3] and Damini Singh Thakur[4]
[1,2,3,4] Dept. of Computer and Information Science, SUNY Polytechnic Institute, Utica, NY, USA
Email: {bharata1@sunypoly.edu, anakals@syr.edu, gadev@sunypoly.edu, thakurd1@sunypoly.edu}

***Abstract*— The abstract of "Mining Future Scopes from Research Manuscripts" would detail the utilization of data mining methods to extract information from research papers that can assist in identifying potential future research paths. It would explore the challenges associated with identifying future research directions from existing literature and the benefits of using data mining techniques to address these challenges. The abstract would also outline the particular data mining techniques employed in the research, such as citation analysis or topic modeling, and provide an overview of the results obtained from their application.**

Scientists have created a method that uses natural language processing to extract potential future research pathways from scientific papers. The method involves identifying important sentences, selecting key phrases, grouping them together based on similarity, and summarizing the most promising research areas. The study tested this approach on a collection of scientific papers and found that it successfully identifies and summarizes essential future directions. This technique can be beneficial for researchers, publishers, and funding agencies to swiftly identify possible research areas and avoid redundant efforts.

***Index Terms*— Topic Modelling, Data Mining, citation analysis.**

Introduction

Scientific research is a continuous endeavor to uncover and enhance knowledge across various fields. As researchers aim to broaden their area of expertise, they propose future research directions. However, the growing volume of scientific literature and potential research avenues can pose a significant challenge for them to stay updated. To address research challenges, it can be advantageous for researchers to extract potential future research areas from existing research papers. By doing so, they can efficiently pinpoint promising areas for further exploration, thereby avoiding duplicative efforts and conserving time and resources. This can be accomplished through the automated summarization of the future research directions endorsed by authors in their papers. In recent times, natural language processing has demonstrated its ability to extract useful insights from vast quantities of unstructured text. In this paper, the authors suggest a novel approach that utilizes natural language processing to identify possible avenues for future research from research papers. Their methodology involves pinpointing pertinent sentences, extricating significant phrases, clustering them based on similarity, and synthesizing the most promising research directions.

Through conducting experiments on scientific manuscripts, the effectiveness of a particular approach was evaluated. The findings demonstrated that this approach is capable of successfully extracting and condensing the suggested future research directions provided by authors. This automated technique has the potential to be a valuable resource for a variety of individuals in the scientific community, including researchers, publishers, and funding agencies. By utilizing this method, they can quickly and easily identify various research areas and opportunities, ultimately accelerating the process of scientific discovery. With the advent of digital archives and repositories for scientific literature, new opportunities have arisen for analyzing scientific manuscripts through various techniques which is used for particular data. As a result, there has been an increasing curiosity in utilizing automated methods to identify potential areas of research for the future from these papers. The process will be advantageous for researchers as it will aid them in saving a significant amount of time by assisting them in selecting research topics. The identification of potential research areas from research papers will become a seamless task using this process.

Motivation & Background

In this section, we first discuss the motivation of our work, and then we briefly discuss some background terms related to our work.

A. Motivation

The primary objective of extracting potential future prospects from research papers is to accelerate the pace of scientific development and avoid duplication of research efforts. By identifying potential breakthroughs and innovations, researchers can focus their efforts on unexplored avenues of research and avoid wasting resources on topics that have already been thoroughly examined. This process can ultimately lead to more rapid discoveries and advancements in various fields of study. Additionally, by staying abreast of the latest research findings and identifying areas of potential growth, researchers can sense what is happening and they can see the latest trends which they can work upon in their fields. Thus, the motivation behind extracting future possibilities from research papers is to maximize the efficiency and effectiveness of scientific research and foster a culture of innovation and progress.

It has become challenging to keep up with the latest research trends and locate valuable research opportunities due to the vast amount of scientific information accessible online. To prevent wasting resources on extensively studied areas, it is essential for researchers, publishers, and funding agencies to promptly recognize promising research directions.

This paper will help lot of people to get idea about what field they are interested in, and they can do the research very rapidly comparatively.

One of the crucial tasks of researchers is to pinpoint promising areas for future research. This serves the purpose of both building upon past research and generating novel ideas. However, such information may not always be easily accessible or concisely summarized in the abstract or conclusion of a paper.

The utilization of automated methods to extract information regarding upcoming research areas from academic papers can prove to be highly beneficial for researchers, publishers, and funding organizations. Such techniques employ machine learning algorithms to scrutinize the language utilized in papers, thereby pinpointing the most promising areas for future research. By doing so, these techniques can provide a concise summary of crucial areas that necessitate further investigation. This expedites scientific progress by enabling researchers to concentrate their efforts on areas that hold the most potential, thereby avoiding wastage of resources and time and ultimately leading to more meaningful research outcomes.

In order to address the difficulties faced in real-time, our team is driven to create a methodology that involves the creation of an algorithm capable of predicting the upcoming research areas. This innovative approach will not only save valuable time and energy for scientific researchers, but also prove to be a valuable asset for anyone interested in pursuing a specific field of research.,

B. Background

This paper utilizes certain techniques and terminology in the creation of its algorithm, which will be discussed briefly in this section.

Topic modeling is a computer-based method used in NLP and machine learning to discover hidden topics or themes in a group of documents. It's an unsupervised learning algorithm that can detect patterns in text data without needing predetermined categories or labels.

Latent Dirichlet Allocation (LDA) is the most popular algorithm for topic modeling. LDA can determine the most likely topics covered by a document by examining the frequency of words used.The use of topic modeling has gained popularity in recent times as digital archives have grown, and there is a need to analyze large amounts of text data automatically. It has found applications in various fields, such as social media analysis, academic research, and customer reviews.

Common uses of topic modeling are that can be used to automatically identify and analyze the main topics in large collections of text data. This technique has various applications in different fields, including classifying documents into categories, identifying trends over time, recommending similar documents or products, and generating new content relevant to the most common topics. Overall, topic modeling is a powerful tool for analyzing latent topics in text data.

DATA COLLECTION

Observations or measurements are systematically gathered through the process of data collecting. Whether you are conducting research for commercial, governmental, or academic goals, data gathering enables you to get first-hand expertise and unique insights into your study challenge. Data collecting is one of the most important components of our research. The results will be more accurate the more accurate the data collection method is.

In our project, we used ASE conferences to gather data from a stream of academic publications on software engineering starting in 2021. From the final section of the research papers, we have gathered the following sentences. Future sentences are typically seen in the conclusion and future work paragraphs of the research paper's last section. From the software engineering research articles from 2021, we gathered 1200 future phrases. As we talked earlier the more the sentences the more accurate results will be. We have collected the future sentences manually from the software engineering research papers from 2021. We have collected the data manually to train the model, so that model can predict the future sentences. We manually collected information from Google Scholar metrics for the software engineering stream. Future sentences, p, q, and r files were created from the information. The data was trained on three different sets of folders. We separated the 1200 sentence dataset into three subsets, corresponding to datasets for sentences, respectively.

DATA PREPROCESSING

The Data can be load into google colab for data preprocessing. Google Colab is a web-based tool that enables people to execute and distribute Jupyter Notebook files, which include code, text, and graphics. This platform offers a free Python programming environment that includes GPUs and TPUs to support machine learning and data analysis duties.

Data scientists and machine learning practitioners prefer Colab because it enables them to handle big data and build sophisticated models without costly equipment. It furthermore offers a collaborative space for users to share and collaborate on projects in real-time. Colab requires users to have a Google account and can be accessed through a web browser. Users can create or upload notebooks and execute code cells individually or as a whole. It comes with pre-installed packages like TensorFlow and PyTorch, which simplifies the process of starting machine learning and data analysis tasks.

Google Colab is a valuable resource for data scientists and machine learning experts as it allows them to run code and analyze data in a free and collaborative environment.

Our project's data is organized using a Jupyter notebook. In this notebook, we will divide the tokens into individual sentences and assign a specific number to each token. The tokens will be stored in a tensor, which can be described using an n-dimensional numerical array. Essentially, a tensor is a generalized matrix. These tokens will be utilized in training the model.

Data preprocessing is a crucial step in data mining that enables the conversion of raw data into a usable and effective format. This process encompasses various tasks, including data cleaning, which is vital in ensuring the accuracy and reliability of the data. Among the tasks involved in data preprocessing are tokenization, stop word removal, punctuation removal, lemmatization, and lowercase processing, which all work together to produce preprocessed data that is ready for analysis.

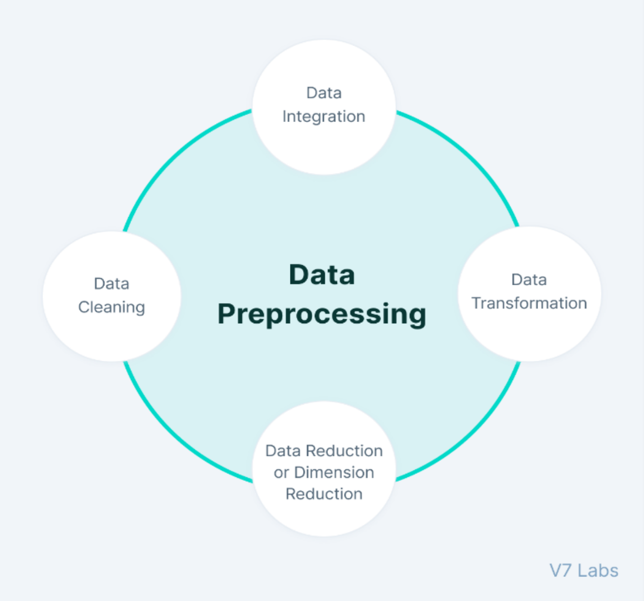


Fig 1. Data preprocessing steps

APPROACH

Latent Dirichlet Allocation (LDA) is a probabilistic model used in topic modeling to identify the underlying topics in a collection of documents. The approach involves analyzing the frequency of words in each document and then identifying the most probable topics that the document covers.

The LDA model assumes that each document that can take as mixture of words and each topic is assigned by a distribution of words. The goal is to identify the distribution of topics across all the documents in the collection, as well as the distribution of words within each topic.

The approach begins by randomly assigning each word in the corpus to a random topic. Then, the algorithm iteratively updates the assignments by considering the probability of each word belonging to each topic and the probability of each topic occurring in each document. This process continues until a stable assignment of words to topics is reached.

The resulting topic model can then be used to identify the most probable topics for each document in the collection, as well as the most probable words for each topic. The model can also be used to generate new documents by randomly selecting words from the learned topic distributions.

Overall, LDA is a powerful approach for automatically identifying the underlying topics in a collection of documents, and has been used in various applications, including text classification, trend analysis, and content generation. However, it also has limitations, such as the need for manual tuning of hyperparameters and the difficulty of interpreting the resulting topics.

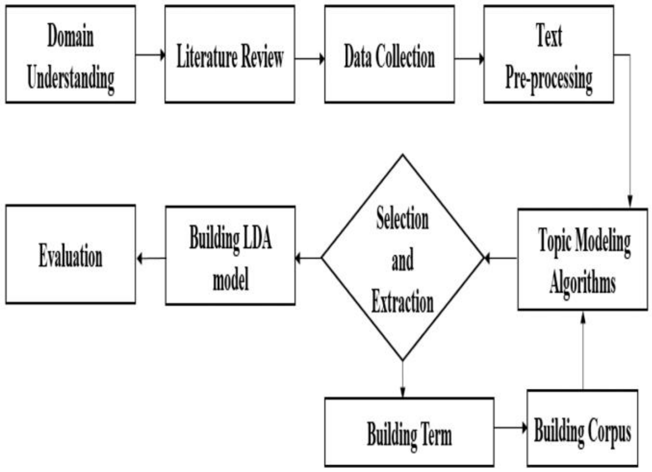


Fig 2. LDA Architecture

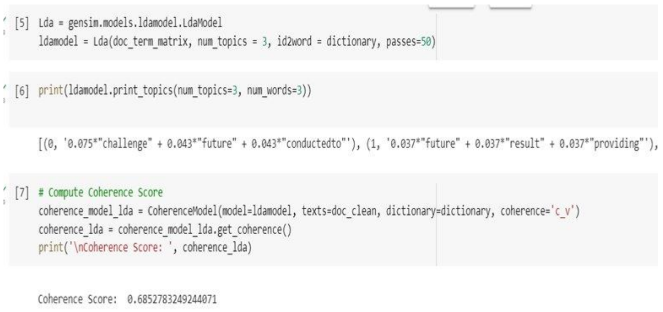
RESULTS

The results are calculated by getting the coherence score.
The coherence score is a tool utilized to assess the effectiveness of topic modeling algorithms, namely LDA, in generating high-quality topics. This metric evaluates the level of interconnectedness between the primary words within a topic and their uniqueness in comparison to the primary words of other various fields.

The coherence score is an important metric in LDA, as it helps to identify the best number of topics to create. Through the comparison of coherence scores across various topic models, researchers can select the model that exhibits the highest coherence score, indicating a set of topics that are both meaningful and interconnected.

The coherence score is a metric typically determined through the use of external resources such as WordNet or Wikipedia, which offer a way to measure the semantic similarity between words. There are multiple techniques for calculating coherence score, such as the pointwise mutual information (PMI) method and the cosine similarity method.

The coherence score serves as an advantageous tool for assessing the effectiveness of topic modeling algorithms in generating high-quality topics. It enables researchers to determine the most relevant and logically connected topics in their data, thereby enhancing their ability to derive meaningful insights from it.



```
[5] Lda = gensim.models.ldamodel.LdaModel
    ldamodel = Lda(doc_term_matrix, num_topics = 3, id2word = dictionary, passes=50)

[6] print(ldamodel.print_topics(num_topics=3, num_words=3))

    [(0, '0.075*"challenge" + 0.043*"future" + 0.043*"conductedto"'), (1, '0.037*"future" + 0.037*"result" + 0.037*"providing"'),

[7] # Compute Coherence Score
    coherence_model_lda = CoherenceModel(model=ldamodel, texts=doc_clean, dictionary=dictionary, coherence='c_v')
    coherence_lda = coherence_model_lda.get_coherence()
    print('\nCoherence Score: ', coherence_lda)

    Coherence Score:  0.6852783249244071
```

RELATED WORK

The related work to our project involves analyzing Amazon reviews. Amazon conducts sentiment analysis on the reviews they have collected and categorizes the output into three categories. The analysis takes into account that some customers give a product a high rating and express their satisfaction, while others give a low rating and express their dissatisfaction. This presents a challenge for companies trying to gauge customer satisfaction and for customers themselves. To address this issue, a new approach to sentiment analysis has been proposed, which predicts the polarity of reviews or ratings using a binary classification problem. The insights gained from Amazon reviews help us to understand how we can tackle similar issues in our project.

The other related work similar to our project is Detecting Health Advice in Medical Research Literature is a project in which discusses the development of a machine learning approach for identifying health advice in medical research literature. The authors present a dataset of medical research articles labeled with health advice statements and use this dataset to train and evaluate different machine learning models. They find that a combination of lexical and syntactic features, as well as domain-specific features such as MeSH terms, can be used to accurately identify health advice statements in medical research literature. They also show that their approach outperforms existing methods for identifying health advice statements. The authors argue that the ability to automatically identify health advice statements in medical research literature can have important implications for clinical practice, as it can help clinicians and researchers to quickly identify relevant health advice and integrate it into their practice or research. They suggest that future work could focus on developing more sophisticated machine learning models that can accurately identify different types of health advice statements and improve the whole performance of the approach.

CONCLUSION

By utilizing the practice of research mining manuscripts, researchers can elegantly and persuasively identify gaps in current knowledge and pinpoint areas where further research is necessary. This will assist researchers in making more astute choices in regard to the research projects they choose to pursue. This project offers a highly valuable instrument to both researchers and policymakers alike, as it aids in the identification of areas that hold potential for future research and development. The task of identifying potential areas of research and innovation from research manuscripts is both challenging and crucial. This can be achieved through the utilization of text mining and natural language processing methods to extract pertinent data from voluminous manuscripts, and also through machine learning and data analytics to uncover patterns and trends within the extracted information. This approach aids researchers and practitioners in identifying promising areas for future research and innovation.

In order to effectively extract potential areas of study from research publications in the future, it is imperative to meticulously strategize and carry out the data gathering process, refine the data to guarantee its accuracy and uniformity, and adopt suitable methodologies to scrutinize and construe the data. Additionally, it is crucial to conduct a comprehensive analysis of pre-existing literature to pinpoint any possible areas for advancement and steer the creation of innovative techniques and methods. By effectively extracting potential opportunities and directions from existing research documents, scientists and professionals have the potential to obtain novel understandings about emerging patterns and research fields. This knowledge can be utilized to enhance their own research and practice, ultimately resulting in the creation of innovative technologies, therapies, and actions that can enhance health results and progress the medical industry.

FUTURE WORK

There are several potential areas for future work in the field of mining future scopes from research manuscripts. Some of these areas include are Most studies in this field only use research manuscripts as their main source of data, but including other sources such as clinical trial data, electronic health records, or social media data could offer more understanding of upcoming trends and research areas. The use of machine learning in identifying future research scopes in medical research manuscripts has potential, but there is still room for improvement. One area for improvement is the development of more advanced models that can better differentiate between different types of future research scopes and enhance the overall effectiveness of the approach. Deep learning techniques, specifically neural networks, have displayed potential in multiple natural language processing tasks. Further research could investigate the application of these techniques in detecting potential areas of research in medical research manuscripts.

Although some research has looked into using domain-specific features like MeSH terms, there is still much to discover about which features are most effective for identifying potential research areas in medical manuscripts. Creating interactive tools for researchers and practitioners can simplify the process of mining future research possibilities from manuscripts. This could promote the use of this method in practice. There is still a lot to discover about the most effective methods for extracting future research possibilities from academic papers. By experimenting with different strategies, scientists and professionals can gain fresh perspectives on upcoming trends and research areas and apply this knowledge to the medical field's progress.

REFERENCES:

[1] Sun, M., Huang, X., Ji, H., Liu, Z., & Liu, Y. (Eds.). (2019). Chinese Computational Linguistics. Lecture Notes in Computer Science. doi:10.1007/978- 3-030-32381-3.

[2] S. Sagnika, A. Pattanaik, B. Shankar, P. Mishra, and S. K. Meher, "A Review on Multi-Lingual Sentiment Analysis by Machine Learning Methods", Journal of Engineering Science and Technology Review, vol. 13, no. 2, pp. 154 – 166, April 2020.

[3] Maarten Grootendorst, 'Keyword Extraction with BERT', Oct 29, 2020. [Online]. Available: https://towardsdatascience.com/keyword-extractionwith-bert-724efca412ea [Accessed on: Nov-20- 2022].

[4] S. A. a. A. N. S. Aljuhani, "A Comparison of Sentiment Analysis Methods on Amazon Reviews of Mobile Phones," International Journal of Advanced Computer Science and Applications, vol. 10, 2019.

[5] L. a. L. B. Zhang, "Aspect and entity extraction for opinion mining," in Zhang, Lei and Liu, Bing, Berlin, Heidelberg, Springer, 2014, pp. 1--40.

[6] Y.-C. a. K. C.-H. a. C. C.-H. Chang, "Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor," International Journal of Information Management, vol. 48, pp. 263--279, 2019.

[7] K. S. a. D. J. a. M. J. Kumar, "Opinion mining and sentiment analysis on online customer review," in IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016.

[8] A. S. a. A. A. a. D. P. Rathor, "Comparative study of machine learning approaches for Amazon reviews," Procedia computer science, vol. 132, pp. 1552--1561, 2018.

[9] B. a. S. S. Bansal, "Sentiment classification of online consumer reviews using word vector representations," Procedia computer science, vol. 132, pp. 1147--1153, 2018.

[10] A. a. S. V. a. M. B. ernian, "Sentiment analysis from product reviews using SentiWordNet as lexical resource," in 2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Bucharest, 2015.

[11] J. F. V. W. P. G. N. Rodrigo Moraes, "Document-level sentiment classification: An empirical comparison between SVM and ANN," Expert Systems with Applications, vol. 40, pp. 621-- 633, 2013.

[12] D. a. X. H. a. S. Z. a. X. Y. Zhang, "Chinese comments sentiment classification based on word2vec and SVMperf," Expert Systems with Applications, vol. 42, pp. 857--1863, 2015.

[13] Y. a. L. M. a. E. K. K. E. Al Amrani, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," Procedia Computer Science, pp. 511-520, 2018.

[14] Y. a. K. V. Saito, "Classifying User Reviews at Sentence and Review Levels Utilizing Naïve Bayes," in 21st International Conference on Advanced Communication

Technology (ICACT), PyeongChang Kwangwoon_Do, Korea (South), 2019.

[15] X. C. T. S. M. W. N. J. Sobia Wassan, "Amazon Product Sentiment Analysis using Machine," Revista Argentina de Clínica Psicológica, pp. 695-703, 2021.

[16] Thet, T.T.; Na, J.; Khoo, C.S.G. Aspect-based sentiment analysis of movie reviews on discussion boards. J. Inf. Sci. 2010, 36, 823–848. [Google Scholar] [CrossRef]

[17] Pota, M.; Ventura, M.; Fujita, H.; Esposito, M. Multilingual evaluation of pre-processing for BERTbased sentiment analysis of tweets. Expert Syst. Appl. 2021, 181, 115119. [Google Scholar] [CrossRef]

[18] Ben. Lutkevich, "BERT language model". techtarget.com. https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model (accessed Nov. 20, 2022).

[19] Y. Jeong and E. Kim, "SciDeBERTa: Learning DeBERTa for Science Technology Documents and Fine-Tuning Information Extraction Tasks," in IEEE Access, vol. 10, pp. 60805-60813, 2022, doi: 10.1109/ACCESS.2022.3180830.

[20] Product Sentiment Analysis for Amazon Reviews. Arwa S. M. AlQahtani "International Journal of Computer Science & Information Technology (IJCSIT) "Vol 13, No 3, June 2021

[21] Sonbol, R., Rebdawi, G., & Ghneim, N. (2022). The Use of NLP-Based Text Representation Techniques to Support Requirement Engineering Tasks: A Systematic Mapping Review. arXiv. https://doi.org/10.1109/ACCESS.2022.3182372.
[22] CUET-NLP@TamilNLP-ACL2022: Multi-Class Textual Emotion Detection from social media using Transformer (Mustakim et al., DravidianLangTech 2022.