University of Calcutta

Semester III

MCA-P33 (Mini-Project)

Document Contextual Similarity Evaluation

Under the supervison of Prof. Amlan Chakraborti

Pritam Sarkar (Roll - C91/MCA/222013)
Akshay Kumar Das (Roll - C91/MCA/222001)

January 31, 2024

Document Contextual Similarity Evaluation

Pritam Sarkar & Akshay Kumar Das

January 31, 2024

Abstract

In the ever-expanding landscape of academic research, ensuring the uniqueness and novelty of one's ideas is a paramount concern. This work addresses the challenge of evaluating document contextual similarity, offering a solution that empowers researchers to gauge the distinctiveness of their contributions. By leveraging advanced transformer model's capability to perform state-of-the-art NLP tasks given a diverse dataset encompassing research, journals, and academic works from various domains, our approach facilitates a nuanced comparison of the contextual meanings embedded within different documents.

Our methodology involves fine-tuning a transformer model on a comprehensive dataset, enabling it to capture intricate patterns and contextual nuances present in academic literature. The model's capabilities extend to assessing the contextual relevance between a given document and a corpus of existing works. A crucial aspect of our approach is the incorporation of a threshold mechanism, allowing researchers to set a similarity threshold. When the similarity score between a researcher's document and those in the repository surpasses this threshold, it indicates potential similarities or overlaps in ideas.

This abstract provides a comprehensive overview of our endeavor to enhance the discernment of research uniqueness through document contextual similarity evaluation. We delve into the intricacies of our methodology, which combines cutting-edge transformer models with a robust dataset, fostering a more informed and insightful research landscape..

1 Introduction

The realm of academic research is a dynamic and collaborative space where ideas continually evolve and interconnect. As researchers embark on exploring new territories, the critical question of whether their ideas align with existing literature emerges as a pivotal concern. To address this, our work introduces a sophisticated framework for evaluating document contextual similarity, offering a nuanced perspective on the interconnectedness of academic contributions.

In this era of information abundance, the sheer volume of research publications poses a challenge to researchers seeking to ascertain the uniqueness of their ideas. The proposed framework seeks to bridge this gap by harnessing the power of transformer models, specifically fine-tuned on a diverse and extensive dataset comprising research articles, journals, and academic works spanning various domains.

Our approach takes inspiration from recent advancements in natural language processing, focusing on the development of a transformer model capable of capturing intricate contextual nuances within documents.

Central to our methodology is the introduction of a threshold mechanism, allowing researchers to customize the sensitivity of the similarity evaluation. By setting a threshold, researchers can define the level of similarity that prompts further investigation. When a researcher submits a document for evaluation, the model computes a similarity score against a diverse set of articles from different domain present in the source dataset or in the repository. If this score surpasses the defined threshold, it suggests potential similarities or overlaps in the contextual meanings of the documents.

In the subsequent sections, we delve into the technical intricacies of our methodology, shedding light on understanding the contextual meaning of articles, model architecture, Our ultimate aim is to empower researchers with a tool that not only evaluates document similarity but also enriches their understanding of the broader academic landscape, fostering collaboration and informed exploration of new research frontiers.

2 Related Work

In the paper "Aspect-Based Document Similarity Using Transformer Models" *Ostendorff et al., 2020* [1] introduces an aspect-based document similarity measure, leveraging Transformer models like SciBERT for enhanced granularity in distinguishing similar documents. The focus on incorporating aspect information into document similarity opens avenues for more nuanced applications, particularly in recommender systems. Finally in the paper "Comparative Evaluation of Document Similarity Algorithms" *Gahman & Elangovan, 2023*[3] focused on identifying the most effective document similarity algorithm, this paper comprehensively evaluates statistical, neural network, and knowledge-based approaches. The findings highlight the superiority of the MT-DNN neural network model across multiple natural language processing tasks, providing valuable insights into algorithm selection for tasks like text summarization and plagiarism detection. Similarly in the paper *Mathur & Joshi, 2012*[4] the document presents a comprehensive overview of a plagiarism detection system addressing the issue of detecting plagiarism in documents. The system utilizes natural language processing and web search technologies to analyze text, extract keywords and chunks, and search for matches online. The two-phase architecture involves processing input documents to extract keywords and chunks, followed by a search for these chunks online using APIs. The system calculates plagiarism percentages and generates a detailed report highlighting copied sentences and their sources.

3 Objective

The primary objective of this project is to develop an advanced system for assessing the contextual similarity between academic documents, providing researchers with a powerful tool to gauge the uniqueness of their contributions within a specific domain. The central focus is on implementing a robust methodology to scrutinize a document's contextual meaning concerning other documents present in the repository. The goal is to assist researchers in determining the novelty of their work by evaluating how closely their ideas align with existing literature. Leveraging a substantial dataset comprising diverse research, journal articles, and academic works from various domains, the project involves fine-tuning transformer models to capture intricate contextual nuances. The key output is a similarity score, computed against a repository of articles from various domain, with a user-defined threshold to alert researchers when their work bears significant resemblance to existing literature. The system aims to empower researchers in ensuring the originality of their contributions and making informed decisions about the novelty of their ideas in the academic landscape. The project will culminate in a comprehensive system and documentation, offering a valuable resource for researchers seeking to assess the uniqueness of their work.

4 Proposed Methodology

1. **Load Dataset:**
   Our process commences with the acquisition of a diverse dataset comprising article from various domains. This dataset forms the basis for the subsequent steps in our methodology.

2. **Pre-process and Optimize Data for Quality Assurance**
   Prior to fine-tuning the transformer model, the acquired dataset undergoes a pre-processing phase. During this phase, data is cleaned, redundant information is removed, and the dataset is optimized to ensure the highest quality possible. Quality assurance measures are implemented to enhance the overall reliability of the source data.

3. **Source Vectors: Convert source articles to Vectors Using Word Embeddings**
   Leveraging the fine-tuned transformer model's word embedding capability, we transform the textual data within the source dataset into numerical vectors. The concept encapsulated in these vectors represents the semantic richness of the source data. This step ensures that the source vectors

not only capture individual words but also the collective meaning and contextual relationships within the entire document.

4. **Incoming Document Vector: Average Embeddings for Incoming Document Representation**
For the incoming document, we employ the fine-tuned transformer model to generate word embeddings. These embeddings are then averaged to create a document vector that encapsulates the overall contextual meaning of the document. This process ensures that the model captures the distinctive features of each document, facilitating a comprehensive representation for subsequent similarity analysis.

5. **Language Handling: Detect and Translate Non-English Articles**
In pursuit of inclusivity, our approach includes a language handling component. Beginning with language detection for each document, we identify non-English articles. Subsequently, a translation process is applied to these articles, converting them into English for consistent and unified analysis. This step ensures that the model evaluates contextual similarity across documents with diverse linguistic backgrounds.

6. **Analysis: Contextual Similarity Between Document and Source Vectors**
The core of our approach lies in evaluating the contextual similarity between the incoming document and the source vectors obtained from the source dataset. Utilizing advanced cosine similarity metrics, we quantitatively measure the alignment of contextual meanings. Cosine similarity, being suitable for vector comparison, serves as a robust technique for generating similarity scores, offering researchers nuanced insights into the thematic resonance between their document and existing literature.

7. **Reporting Results: Output Similarity Score**
The final step involves presenting the results of the contextual similarity analysis. Researchers receive an output comprising similarity scores, offering insights into the degree of alignment between their document and existing literature. Additionally, a decision threshold enables researchers to customize the sensitivity of the evaluation, allowing them to tailor the analysis to their specific research needs.

5 Experimental Results

In this section, we present the results of our comprehensive experimentation on the "Document Contextual Similarity Evaluation" project. We are using Covid 19 article dataset as a primary dataset of our experiment. The main objective of the experiments was to assess the effectiveness of various Transformer models in vectorizing Covid-19 article abstracts and evaluating contextual similarity

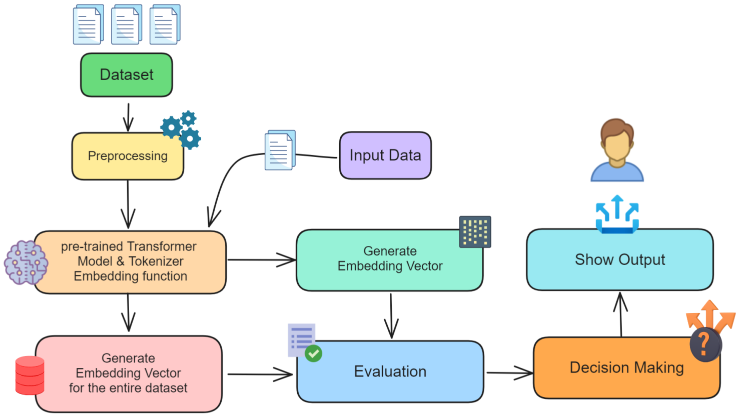

Dataset
Preprocessing
Input Data
pre-trained Transformer Model & Tokenizer Embedding function
Generate Embedding Vector
Show Output
Generate Embedding Vector for the entire dataset
Evaluation
Decision Making

Figure 1: Representation of the proposed methodology

using cosine similarity metrics.

$$(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{1}$$

In this experiment, we are using various transformer models like BERT, SciBERT, RoBERT, and XLNet to check which one gives the best result. And finally, we integrated the streamlit module to get a good user interface. We are choosing these four transformers because BERT Captures intricate word relationships for robust similarity assessments, where SciBERT model is tailored for scientific literature, excelling in research-based text analysis, the RoBERT model is optimized for enhanced training and accurate plagiarism detection, and lastly, XLNet model which adds a dynamic perspective, considering the entire document context.

Firstly we are creating the vector embedding for source dataset article abstracts using all four transformer models we are selecting and storing that embedding as different pickle files. Then we test the input text vector embedding which is generated by the same one of the four transformer models and check the contextual similarity concerning that same stored transformer generated source embedding. Also if the input is in a different language we translate that input to English generating the vectors and performing evaluation. First, we are evaluating the similarity score between the same article abstract and some para-phrased and rewritten versions of the same article abstract present in the source dataset to check whether it can detect the contextual similarity between the same articles or not. Then we evaluate the similarity score of different article abstracts or writing with the source article, in this process, we are checking whether our system can identify the less contextual similarity of both articles or not. The result score of this process is shown below.

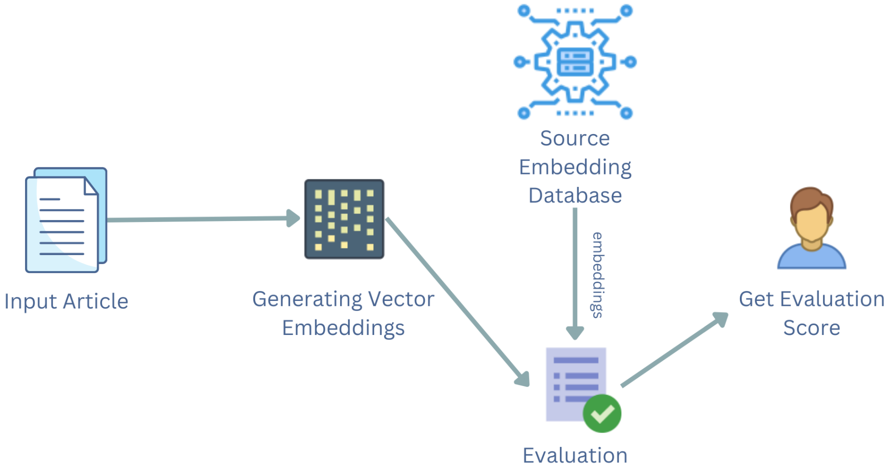
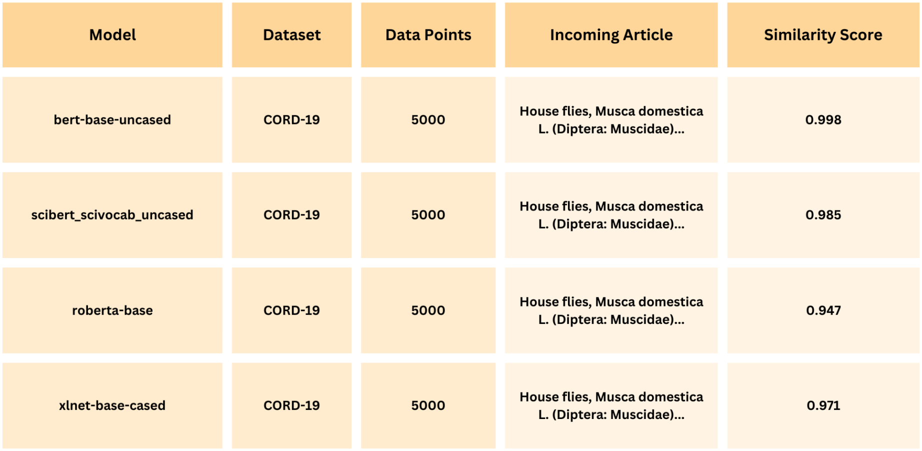
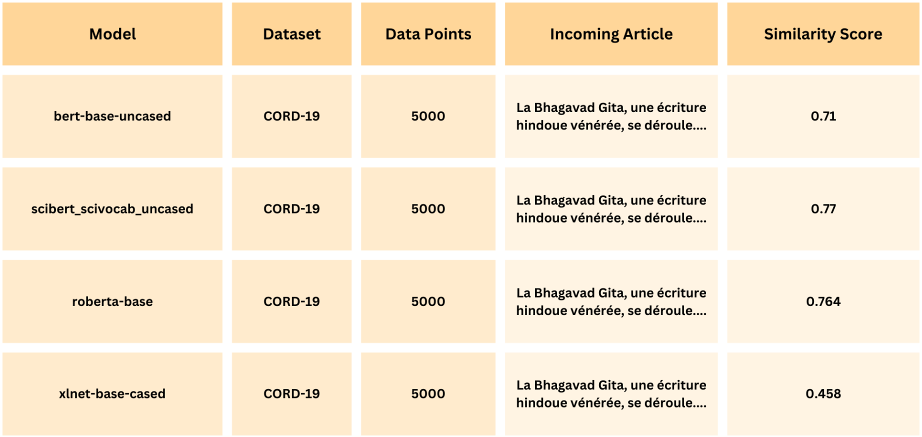


Figure 2: Experiment Process



Model	Dataset	Data Points	Incoming Article	Similarity Score
bert-base-uncased	CORD-19	5000	House flies, Musca domestica L. (Diptera: Muscidae)...	0.998
scibert_scivocab_uncased	CORD-19	5000	House flies, Musca domestica L. (Diptera: Muscidae)...	0.985
roberta-base	CORD-19	5000	House flies, Musca domestica L. (Diptera: Muscidae)...	0.947
xlnet-base-cased	CORD-19	5000	House flies, Musca domestica L. (Diptera: Muscidae)...	0.971

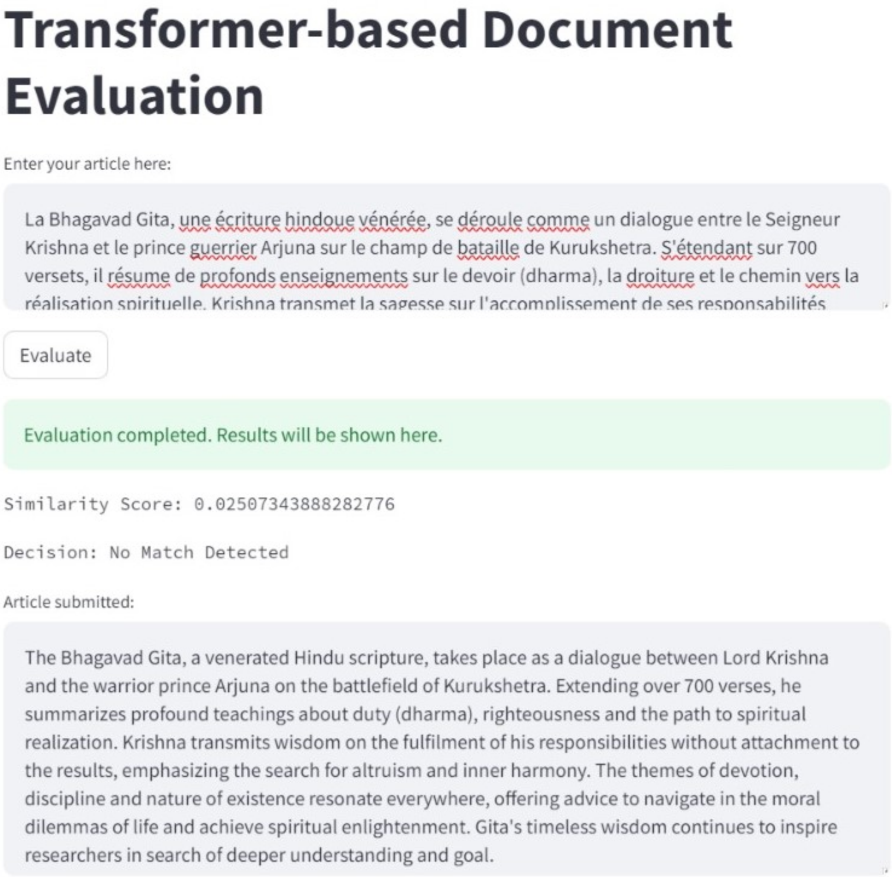
Figure 3: Experiment with similar document writing and the score respective to each transformer model



Model	Dataset	Data Points	Incoming Article	Similarity Score
bert-base-uncased	CORD-19	5000	La Bhagavad Gita, une écriture hindoue vénérée, se déroule....	0.71
scibert_scivocab_uncased	CORD-19	5000	La Bhagavad Gita, une écriture hindoue vénérée, se déroule....	0.77
roberta-base	CORD-19	5000	La Bhagavad Gita, une écriture hindoue vénérée, se déroule....	0.764
xlnet-base-cased	CORD-19	5000	La Bhagavad Gita, une écriture hindoue vénérée, se déroule....	0.458

Figure 4: Experiment with dissimilar document writing and the score respective to each transformer model

As We can see that with respect to similar document each model gives a good similarity score but BERT models score is better then other models score (Figure 3)., but in contrast with dissimilar documents the XLNet models score is much better in compare to BERT or any other transformer models (Figure 4).



Transformer-based Document Evaluation

Enter your article here:

La Bhagavad Gita, une écriture hindoue vénérée, se déroule comme un dialogue entre le Seigneur Krishna et le prince guerrier Arjuna sur le champ de bataille de Kurukshetra. S'étendant sur 700 versets, il résume de profonds enseignements sur le devoir (dharma), la droiture et le chemin vers la réalisation spirituelle. Krishna transmet la sagesse sur l'accomplissement de ses responsabilités

Evaluation completed. Results will be shown here.

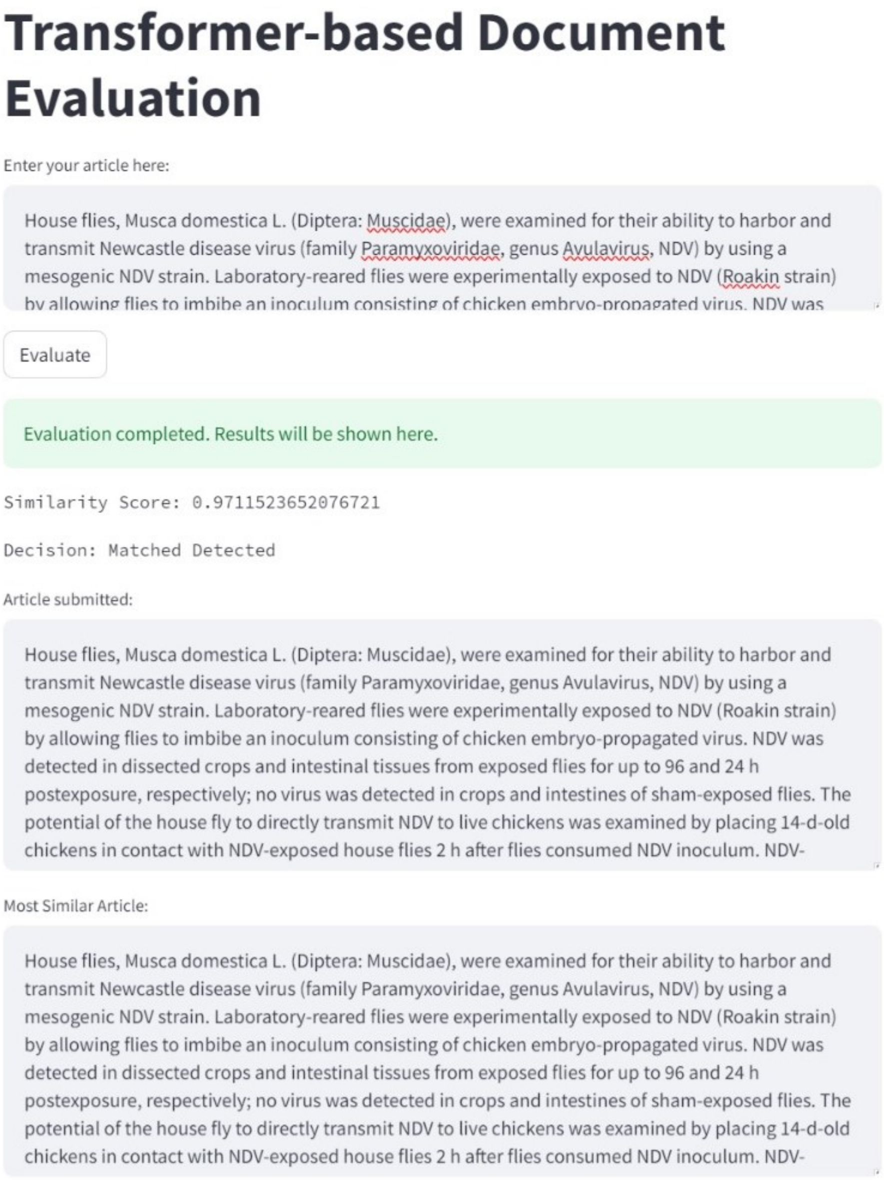
```
Similarity Score: 0.02507343888282776

Decision: No Match Detected
```

Article submitted:

The Bhagavad Gita, a venerated Hindu scripture, takes place as a dialogue between Lord Krishna and the warrior prince Arjuna on the battlefield of Kurukshetra. Extending over 700 verses, he summarizes profound teachings about duty (dharma), righteousness and the path to spiritual realization. Krishna transmits wisdom on the fulfilment of his responsibilities without attachment to the results, emphasizing the search for altruism and inner harmony. The themes of devotion, discipline and nature of existence resonate everywhere, offering advice to navigate in the moral dilemmas of life and achieve spiritual enlightenment. Gita's timeless wisdom continues to inspire researchers in search of deeper understanding and goal.

Figure 5: Streamlit user interface with dissimilar article as an input



Transformer-based Document Evaluation

Enter your article here:

House flies, Musca domestica L. (Diptera: Muscidae), were examined for their ability to harbor and transmit Newcastle disease virus (family Paramyxoviridae, genus Avulavirus, NDV) by using a mesogenic NDV strain. Laboratory-reared flies were experimentally exposed to NDV (Roakin strain) by allowing flies to imbibe an inoculum consisting of chicken embryo-propagated virus. NDV was

Evaluation completed. Results will be shown here.

```
Similarity Score: 0.9711523652076721

Decision: Matched Detected
```

Article submitted:

House flies, Musca domestica L. (Diptera: Muscidae), were examined for their ability to harbor and transmit Newcastle disease virus (family Paramyxoviridae, genus Avulavirus, NDV) by using a mesogenic NDV strain. Laboratory-reared flies were experimentally exposed to NDV (Roakin strain) by allowing flies to imbibe an inoculum consisting of chicken embryo-propagated virus. NDV was detected in dissected crops and intestinal tissues from exposed flies for up to 96 and 24 h postexposure, respectively; no virus was detected in crops and intestines of sham-exposed flies. The potential of the house fly to directly transmit NDV to live chickens was examined by placing 14-d-old chickens in contact with NDV-exposed house flies 2 h after flies consumed NDV inoculum. NDV-

Most Similar Article:

House flies, Musca domestica L. (Diptera: Muscidae), were examined for their ability to harbor and transmit Newcastle disease virus (family Paramyxoviridae, genus Avulavirus, NDV) by using a mesogenic NDV strain. Laboratory-reared flies were experimentally exposed to NDV (Roakin strain) by allowing flies to imbibe an inoculum consisting of chicken embryo-propagated virus. NDV was detected in dissected crops and intestinal tissues from exposed flies for up to 96 and 24 h postexposure, respectively; no virus was detected in crops and intestines of sham-exposed flies. The potential of the house fly to directly transmit NDV to live chickens was examined by placing 14-d-old chickens in contact with NDV-exposed house flies 2 h after flies consumed NDV inoculum. NDV-

Figure 6: Streamlit user interface with similar article as an input

6 Future Work

Moving forward, our focus lies in enhancing the scalability of our model to accommodate larger datasets, addressing current computational constraints through optimization strategies. Fine-tuning efforts will be directed towards tailoring the Transformer models for domain-specific Covid-19 data, ensuring increased relevance. Exploration of advanced NLP techniques, including domain-specific pre-training and attention mechanisms, is paramount to improving accuracy. Overcoming current limitations in computational resources will involve parallelization and optimization, while actively seeking or curating a larger and diverse dataset will contribute to more robust model training. Extensive benchmarking and evaluation against existing methods, along with refinement of metrics, are integral steps. Ultimately, real-world deployment considerations will take precedence, involving integration, runtime efficiency, and usability in practical scenarios. These combined efforts aim to position our "Document Contextual Similarity Evaluation" project at the forefront of advancements in NLP, scalability, and domain-specific applications.

7 Conclusion

In this project, we addressed the critical task of document contextual similarity evaluation, aiming to provide researchers and authors with a tool to assess the uniqueness and novelty of their work. The primary objective was to determine how closely an incoming document aligns with existing knowledge in a given domain, facilitating the identification of potential overlaps or similarities.

Our methodology involved retrieving article abstracts, converting them into source vectors using word embeddings, and generating document vectors for incoming articles. To ensure quality assurance, we pre-processed and optimized the data, considering the language aspect by detecting and translating non-English articles. The analysis was performed using cosine similarity, a suitable metric for comparing vectors, providing a quantitative measure of contextual similarity.

The choice of transformer models, including BERT, sciBERT, RoBERTa, and XLNet, allowed us to explore different pre-trained architectures, with XLNet demonstrating superior performance in the majority of cases. The transformer models were fine-tuned to capture contextual nuances accurately, ensuring a nuanced understanding of document similarities.

The experimental results highlighted the effectiveness of our approach in identifying contextual similarities across a diverse set of documents, both known and unknown. The system's ability to provide similarity scores enables researchers to gauge the originality of their work and identify related literature comprehensively.

In conclusion, our document contextual similarity evaluation system offers a valuable contribution to the academic and research community, empowering authors with insights into the uniqueness of their contributions. As the landscape of knowledge continues to expand, tools like ours become increasingly essential for researchers to navigate and contribute meaningfully to their respective fields.

8 Acknowledgement

We extend our sincere gratitude to Prof. Amlan Chakraborti, Director of the AKCSIT Department, University of Calcutta, for his invaluable patience and constructive feedback. Additionally, we express our appreciation to Sir Wazib Ansar, Research Scholar at the University of Calcutta, for his valuable knowledge and unwavering support.

The guidance and insights provided by Prof. Chakraborti and Sir Ansar were instrumental in shaping the trajectory of our work. Their commitment to excellence significantly enriched our understanding and contributed to the refinement of this technical report. We also acknowledge the combined expertise and assistance of our team members, whose collaborative effort played a vital role in the successful completion of this project.

References

[1] Ostendorff, M., Ruas, T., Blume, T., Gipp, B., & Rehm, G. (2020). ”Aspect-based Document Similarity for Research Papers.” arXiv preprint arXiv:2010.06395.

[2] Vor der Brück, T., & Pouly, M. (2024). ”Estimating Text Similarity based on Semantic Concept Embeddings.” arXiv preprint arXiv:2401.04422.

[3] Gahman, N., & Elangovan, V. (2023). ”A Comparison of Document Similarity Algorithms.” arXiv preprint arXiv:2304.01330.

[4] Mathur, I., & Joshi, N. (2012). ”Plagiarism Detection: Keeping Check on Misuse of Intellectual Property.” arXiv preprint arXiv:1210.7678.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv Preprint arXiv:1810.04805v2.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). ”Attention is All you Need.”