Journal of Information Systems Engineering and Management
2025, 10(18s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Parallelized Influenced Citation Analysis for Medical Documents using Modified BioBERT or ClinicalBERT Model Semantic Similarity Measure

Majji Venkata Kishore[1], Prajna Bodapati[2]
1Research Scholar, Department of CS & SE, AUCE(A) at Andhra University, Visakhapatnam, Andhra Pradesh, India
2Professor, Department of CS & SE, AUCE(A) at Andhra University, Visakhapatnam, Andhra Pradesh, India

ARTICLE INFO

Received: 15 Dec 2024

Revised: 26 Jan 2025

Accepted: 15 Feb 2025

ABSTRACT

Doctors have a limited amount of time, which is quite valuable. They can be wasting their time by reading a variety of publications that have the same semantic substance. Since it emphasizes the importance of the study, gives credit to the researchers, and helps prevent semantic repetition, citation analysis is a crucial part of medical research,and assists in avoiding semantic repetition. Citations are often used by researchers for a variety of reasons, but the primary purpose has to do with demonstrating how the references have impacted their work. These days, the majority of references have a tendency to have some kind of impact on research. The suggested technique takes a semantic approach rather than depending just on keyword matching in order to discover impacted citations in medical research articles. This is accomplished by eliminating the need for keyword matching. References or a bibliography are often included at the conclusion of each and every publication that pertains to medical study. The purpose of this model is to determine, within a semantic framework, the degree of relevance that exists between the medical research document and the list of reference papers that it contains or among given research papers.

In this study, the evolution of the suggested work is described in two different ways: the first is the introduction of a new semantic similarity measure that is used to compute impacted citation scores. This measure is calculated by using Modified BioBERT or ClinicalBERT. Up to this point, every single semantic metric has been calculated by means of using concurrent execution. We are doing this by leveraging parallel algorithms in order to make the process of detecting semantic similarity more efficient. In the end, the work that is being offered is able to efficiently identify affected citations in medical research publications by using high-performance computing (HPC).

Keywords: ClinicalBERT,HPC,Semantic similarity,Modified MWD,Parrallel Computing

1. INTRODUCTION

Citation-based data can be used for a number of purposes, such as assessing the standing of specific academics and institutions. It is normal practice to use standard citation metrics in order to evaluate the influence that scholars have had. Having a solid understanding of the function that references play inside a text is essential for performing fruitful research. Citations inside a paper contribute to the strengthening of its arguments and the establishment of intellectual linkages especially in medical research papers. Citations to a paper, on the other hand, enable communities to evaluate the intellectual contributions and overall quality of the medical research article[1].

An investigation on the significance of reference papers in relation to the primary paper is carried out in this study. Although keyword-based analysis is often used to evaluate relevance, the suggested model largely employs a semantic-based method to evalute the similarity score between reference articles and the primary publication. This is in contrast to the general practice of using keyword-based analysis. This paradigm involves the introduction of a document semantic matching corpus that has thorough annotations[2]. This corpus has the potential to serve as the

ground truth for assessing semantic matching at the document level. Text semantic matching is used extensively in a variety of domains, such as machine translation, automated question answering, and knowledge retrieval, among others. In the academic sphere, it also plays an important part in the detection of plagiarism, the automation of technical surveys, the recommendation of citations, and the study of research trends[3]. The field of text semantics, which encompasses both word semantics and sentence semantics, has been receiving an increasing amount of attention over the last several years[4]. On the other hand, owing to the intrinsic difficulty of document-level semantic matching, there is a limited amount of study on the subject. Long documents often have complex structures and enormous volumes of information, which makes it difficult to evaluate the semantic similarity between them. As far as we are aware, there isn't currently a publicly available dataset created especially for this use[5].

Multiple smaller text units are used to construct a larger text. Through the process of merging the meanings of these smaller components, it is possible to comprehend the meaning of a lengthy text. This method has been used in a great number of recent research in order to ascertain the degree of semantic similarity that exists across bigger text chunks. With the help of the integration of the semantic similarities that exist between word pairs in two different phrases, it is possible to estimate the semantic similarity score of a sentence[6].

There is a substantial lack of research that particularly investigates the semantic similarity between different texts. In lengthy texts, there are often several subject changes and a variety of emphasis points, which makes it difficult to understand the content of the document in its entirety[7].

Background

Building models that learn effective representations of clinical material is difficult. Clinical text has been modeled using bag-of-words assumptions, as well as log-bilinear word embedding models like Word2Vec[8]. The latter word embedding models learn clinical text representations based on local word contexts. However, because clinical notes are long and contain interdependent words, these techniques cannot capture the long-range connections required to capture clinical meaning[9].

Natural language processing approaches that use global, long-range information can improve performance on clinical tasks. Modeling clinical notes necessitates capturing interactions among distant terms. Because of the necessity to depict this long-range structure[9], clinical notes lend themselves to contextual representations such as bidirectional encoder representations from transformers (bert). Apply Bert to biomedical literature, utilize Bert to strengthen clinical concepts[10][11].

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a domain-specific language representation model that has been trained on large biomedical datasets. Because of its nearly identical design across tasks, BioBERT performs better than BERT and earlier state-of-the-art models on a variety of biomedical text mining tasks when pre-trained on biomedical corpora[12]. While BERT performs similarly to prior state-of-the-art models, BioBERT outperforms them on three key biomedical text mining tasks: biomedical named entity recognition, biomedical relation extraction, and biomedical question answering[13][14].

ClinicalBERT develops deep representations of clinical text. These representations can provide clinical insights (such as disease forecasts), identify treatment-outcome correlations, and generate corpus summaries ClinicalBERT is a Bert model adaption that tackles the issues of clinical text in clinical corpora. Medical notes are used to teach representations, which are subsequently processed for therapeutic tasks[14][15].

A survey of the literature

The broad literature evaluation focuses on citation analysis, specifically with respect to semantic similarity in a variety of scientific publications, as well as the references that are mentioned in these works. The various research suggest that semantic similarity techniques have been utilized in internet-related applications, particularly for academic objectives such as the detection of plagiarism, the analysis of social media, and the extraction of root terms for the depiction of inter-object associations. The amount of study that has been done on determining the degree of similarity between paragraphs or papers is substantial; however, the amount of work that has been done on determining the degree of similarity between phrases or smaller texts is relatively less. Corpus-based methods, hybrid methods, descriptive feature information-based methods, and word co-occurrence/vector-based document model approaches are the four basic categories of research.

Both hierarchical and non-hierarchical presentation styles of ontology structures have been used in order to conduct an analysis of the theoretical measures of similarity. There is a gap in the way that link and text observations may be employed to develop related measures in accordance with semantic similarity, and Maguitman et al. suggested a method that overcomes this gap. According to the information-theoretic similarity theory, the degree of semantic similarity between two ideas is determined by the amount to which the concepts share meanings as well as the specific meanings connected to each concept. In Information Retrieval (IR) systems, vector-based approaches are frequently used to evaluate similarity. These approaches involve identifying the documents that are most pertinent to a particular query by representing each document as a word vector[16]. This allows queries to be matched with related documents in the collection using a similarity metric.

Mihalcea and colleagues proposed that by evaluating the similarities between the component terms, a combined technique might be utilized to determine the semantic similarity of information. They used two corpus-based measures: LSA (Latent Semantic Analysis) and PMI-IR (Pointwise Mutual Information and Information Retrieval). Six knowledge-based metrics for word semantic similarity were also used. They demonstrated how these approaches may be applied to create a text-to-text similarity metric by merging this data. Using a challenge that required them to recognize paraphrases, they checked their technique[17]. However, a big disadvantage of this methodology is that it estimates word similarity using eight distinct approaches, which may be computationally costly. This is a huge negative.

Li et al. introduced a hybrid technique that assesses the level of text similarity between the two texts by looking at both the syntactic and semantic information found in the texts under comparison. By using their approach, a dynamic joint word set that includes all of the unique terms from both phrases is produced. The WordNet lexical database is utilized to build an initial semantic vector for every sentence. Then, using the two order vectors as the foundation for the computation, an order similarity is calculated. In the end, semantic similarity and order similarity are combined to determine the overall sentence similarity. Feature-based techniques are those that attempt to describe a phrase by using a set of characteristics that have been predetermined[18].

Sahami and colleagues introduced a method for calculating the level of semantic similarity between two requests. This method utilizes small sets of text samples that a search engine provides. This new methodology eliminates issues found in other approaches, like the Cosine similarity coefficient, when assessing the degree of similarity across short texts. The definition of the semantic similarity between two questions is the inner product of the centroid vectors corresponding to those two queries; however, the researchers did not compare their similarity measure with similarity measures based on taxonomy[19].

Chen and associates created a double-checking technique that evaluates the degree of semantic similarity between terms using phrase snippets from internet search engines. Additionally, a number of related metrics were provided, such as the overlap ratio, occurrence double-checking, Jaccard, Dice, and cosine[20].

A technique for calculating document similarity was developed by Hung Chim and Xiaotie Deng. The approach focuses on sentence-based document similarity and uses the Suffix Tree Document (STD) model to analyze pairwise similarities across texts. Web-based metrics that evaluate semantic similarity between words or concepts were developed by Elias Iosif and Alexandros Potamianos. The authors compared the findings of their metrics to those of the most advanced techniques currently available.

Approaches for Story Link Detection were introduced by Francine Chen and colleagues. These approaches evaluate whether two tales are related to the same events or relationships, with the cosine similarity measure serving as the primary instrument of comparison. Through the use of a variety of similarity measures and the utilization of particular joint information from source pairs, this study presents a method that may improve the performance of link identification. Cosine, Hellinger, Tanimoto, and Clarity are some of the various similarity approaches that have been investigated, both on their own and in conjunction with one another. In their discussion of methods for learning semantic similarity between texts, Jaz and colleagues spoke about two different approaches: one way focuses on document similarity, while another method makes use of co-occurrence information[20].

A technique for assessing the degree of similarity between phrases using web pages was proposed by authors Sheetal A. and associates. Finding semantic connections requires the use of semantic similarity techniques. The degree of semantic similarity between sentences is assessed using web-based metrics in these methods, which then compare the results to the most sophisticated approaches currently in use. Five different association computations for

information retrieval that call for traditional matching are presented in this study. These calculations serve as the foundation for the ensuing similarity metrics[21].

For the purpose of evaluating the efficacy of similarity metrics in partitional clustering for collections of text documents, Anna Huang suggested a technique. This approach made use of the basic K-means algorithm and published results on a variety of text document datasets. Additionally, it employed five distance feature-based similarity algorithms that are typically applied in text clustering. When it comes to retrieval of information across languages, Hsun and Yau have developed a method that makes use of semantic similarity measurements. For the purpose of representing documents, they used fuzzy models, and for the purpose of information extraction, they utilized similarity methods. In the field of computing document similarity, there are a substantial number of papers that are included in the literature. The number of studies that concentrate on brief texts, such as those conducted by Aminul Islam and Diana Inkpen, is, nevertheless, rather low[22]. A technique known as Pointwise Mutual Information is used by them in order to determine the significance of nearby words for two target terms. The words that occur often in both lists are taken into consideration, and the PMI values of those words from the other list are merged in order to establish the degree of semantic similarity between the two lists. By using the meanings that are obtained from the similarity of component terms, a hybrid technique that merges corpus-based and knowledge-based methods has been presented for the purpose of evaluating the semantic similarity of information content.

A detailed study of the many different techniques to semantic similarity was presented by Gomaa and Fahmy [23]. The term-based, knowledge-based, and corpus-based methods to semantic similarities were the three methodologies that were investigated by the researchers. Particular attention is paid by the string-based similarity approaches to the sequence of strings and the combinations of characters that are included within the information. On the other hand, the corpus-based method evaluates nine distinct algorithms in order to determine the degree of semantic similarity between words. This is accomplished by examining data that has been acquired from text collections or huge corpora[24].

Paul Vitanyi presented a technique that does not depend on any factors throughout the process of finding the semantic similarity distance between two words. The fields of image recognition, data mining, and similarity retrieval are all potentially useful uses of this method. In his discussion of compression-based similarity, he went into further detail, elaborating on topics such as feature-based similarity, normalized compression distance, knowledge googling, and Google-based similarity[25].

In order to establish a system for measuring semantic similarity, Mohamed Ali Hadj Taieb and his colleagues included hyponyms and depth factors into their methodology. In order to develop a hyponyms network, they made use of the distribution of depth values, which is an approach that is basically founded on ontology principles[26].

An innovative method for computing similarity that is based on characteristics was presented by Yuncheng Jiang and colleagues of the institution. Utilizing Wikipedia, this approach determines the degree of similarity between two ideas, with the concepts being officially provided prior to the assessment of similarity. The models are constructed on the basis of the facts that are accessible in the ontology, and according to Tversky's similarity theory, the degree of similarity between concepts is evaluated based on the attributes that they possess. This technique solves the constraints of prior semantic similarity models that employed Wikipedia. Specifically, this approach tackles the drawbacks[27].

Utilizing edge numbering and information content theory, Jian-Bo Gaowe and his colleagues have developed a novel approach to determining the degree of semantic similarity between two sets of words. By taking into account the shortest weighted route length between matched ideas, this novel technique determines the degree of semantic similarity between the concepts. They simplified the weighted pathways between ideas and established techniques to apply weights to the nodes that link these paths, which allowed them to speed the process of finding similarity based on the similarities between concepts.

A novel approach to determining semantic similarity was developed by Montserrat Batet and her colleagues. This approach makes use of information material derived from a number of different ontologies. In accordance with their understanding of taxonomy, they identify idea pairings. The performance of their system was examined with the use of a biometric dataset and an ontology, and the results showed that there were considerable increases in the calculation of similarity. Each of the three types of ontology-based measures—edge, feature, and information content-

based—has its own approach for assessing the degree of similarity between word and idea pairings. Ontology-based measures may be classified into these three categories[26][27].

During their presentation, Danushka Bollegala and his associates showed how to use search engines to assess a term's semantic similarity. In order to ascertain semantic similarity, they examined snippets as well as the number of pages that were returned by online searches. The snippets of text provide an important background for reading and comprehending the words. An approach known as shallow pattern retrieval was used by them in order to discover the semantic connections that existed between the words that were identified in these excerpts. Using the lexico patterns that were retrieved from the samples, their innovative similarity metric assesses the degree of similarity between the two[26][28].

When it comes to the use of biometrics in genetics, Francisco et al. have adopted a semantic similarity method. They have made two important contributions to the discussion. For the first, it is based on the knowledge of relationships among Gene Ontology words, which demonstrates that the composition of protein groups serves as a good foundation for assessing the similarity between GO terms[28].

2. METHOD

An exhaustive analysis of the existing literature served as the impetus for the creation of a new model with the purpose of recognizing citations in medical research articles that are unethical or caused by influence. The purpose of this model is to carry out certain activities in order to ascertain whether or not a scientific publication and the references that it cites are sufficiently relevant to one another. It takes medical research papers as input and extracts keywords by concurrently deleting stop words from both the primary text and its reference documents using a process known as term elimination. The ontological approaches are then used to these keywords in order to determine the semantic importance of each keyword that is present in the text.

The model that has been presented produces two options one is “a collection of base papers, together with the reference papers or numerous medical publications that relate to those base papers” and second is “calculate a semantic similarity among multiple medical research papers pair wise”.

Tokenization is carried out during the pre-processing stage, and any words that are not essential are eliminated simultaneously. Due to the fact that previous approaches often evaluate similarity based on word frequency, the primary emphasis of this analysis is intended to be on sentence semantic similarity. On the other hand, the word frequency technique often fails to detect semantic similarity, particularly in situations when synonyms are put together in phrases. With the help of this study, a new semantic similarity measure has been developed that is capable of accurately determining the degree of semantic similarity between medical materials. An electronic lexical database is used to hold the semantic distance between words, which is then used to determine the distance between texts. This metric is based on the semantic distance between words. Following the establishment of the semantic distance, the suggested approach computes the similarity of the documents by using a many-to-many matching between the terms.

The algorithm:

For the purpose of computing semantic similarity, the approach described above uses cosine similarity, which has a number of limitations. Both BioBERT and ClinicalBERT use this method.

1. The application will take the following steps:

1. Load BioBERT or ClinicalBERT models from Hugging Face converters.

2. Use pdfplumber to extract text from PDF files.

3. Tokenize the text using a suitable tokenizer.

4. Create embeddings for each PDF document using the model.

5. Determine cosine similarity between embeddings of each pair of documents.
6. Create a similarity matrix to record pairwise similarity scores.

Comparison of the Cosine with Word movers distance:

Advantages

In terms of efficiency, the cosine similarity algorithm is excellent for use with big datasets since it is computationally efficient. It has outstanding performance characteristics when used on high-dimensional vector spaces, like those produced by word embeddings or term frequency-inverse document frequency (TF-IDF).

• Easy to Implement: The cosine similarity value is straightforward to compute and straightforward to put into practice. Determining the angle that exists between two vectors is a task that may be accomplished in a short amount of time, even when dealing with a huge number of documents.

• Cosine similarity is not impacted by the length of the documents since it quantifies the angle that exists between two vectors. This means that it is not sensitive to the length of the documents. This makes it a valid statistic for comparing documents of varied sizes, such as a long paper and a quick abstract, which are both examples of written documents.

• Cosine similarity is useful for sparse vectors, such as those produced by bag-of-words (BoW) or TF-IDF, which are often employed in classic document comparison tasks. This is because cosine similarity is effective for sparse vectors.

Disadvantages:

There is a limited understanding of semantics since the cosine similarity measure does not take into account the semantic similarity of words. For example, two texts that utilize synonyms (for example, "physician" and "doctor") could be seen as being less similar than they really are. The alignment of document vectors is the only thing that it measures; it does not comprehend the meaning of the words whatsoever.

The cosine similarity algorithm does not take into consideration the sequence in which words appear in a text. This is one of the limitations of the method. There is a possibility that the semantics of two papers that include the same words in different sequences will be different; nonetheless, the cosine similarity score will be the same for both documents.

• Poor Performance on Short Texts: When it comes to short texts, especially in situations where word selections are crucial (such as in medical or legal settings), cosine similarity may have difficulty capturing tiny changes in meaning across documents.

Due to the above disadvantages **BioBERT or ClinicalBERT** model semantic measure is not upto the mark.so we proposed a modified **BioBERT or ClinicalBERT** model which uses word movers distance rather than cosine similarity**.** We adjusted the methods described above by utilizing word mover's distance rather than cosine similarity in order to mitigate the problems that were being experienced.

Modified BioBERT or ClinicalBERT:

Traditional methods of determining similarity sometimes fail when two sentences do not have any terms in common, even if the unusual words convey meanings that are comparable to one another. Taking use of word similarity inside the word embedding space, the model that has been given provides a solution to this problem. This brand new similarity metric is known as Modified Word Mover's Distance (MWMD), and it was developed by Microsoft.

Word Mover's Distance (WMD) is based on the idea of optimum transit between word embeddings. It quantifies the effort required to "move" the words of one text to match another, using pre-trained word embeddings as a semantic metric.

The model makes use of the word embeddings from two distinct texts in order to compute the least distance that separates them. This distance is calculated so that one text may transit the semantic space in order to reach the other text. This method cuts down on the amount of time needed for calculation and improves the effectiveness of determining the degree of semantic similarity between test texts. In addition, the model is centered on the closest distance, but it does not take into consideration the transformation of many words into a single word.

Advantages:

- **Semantic Awareness**: WMD captures semantic links between words using pre-trained word embeddings, such as Word2Vec and BERT. It goes beyond surface-level word matching by considering how semantically similar two words are. For example, WMD detects that "doctor" and "physician" are related despite being different words.

• Contextual Similarity: WMD is effective for documents with distinct words but semantically linked. It calculates how far one text's word distribution can "travel" to match another document in the semantic space.

• WMD is sensitive to word order, as embeddings in context differ from those in isolation. For example, the word "bank" in "river bank" will have a different embedding than in "financial bank," and WMD will change accordingly.
• WMD works best for short and complex texts with significant semantic variations, like as medical papers or news stories with sophisticated language. It is adept at detecting minor variations in meaning.

The following is how the model that has been suggested operates:

Modified BioBERT/ClinicalBERT

1. Preprocessing and Model Initialization

We begin by importing the necessary libraries for handling PDF documents, word embeddings, and distance computations. Specifically, we utilize the pdfplumber library for extracting text from PDF files, the transformers library for loading the BioBERT or ClinicalBERT models, and gensim for computing Word Mover's Distance (WMD). Additionally, we download and configure required Natural Language Toolkit (NLTK) resources for tokenization.

Model Selection: The user is prompted to choose between BioBERT or ClinicalBERT, both pretrained on biomedical data, ensuring domain-relevant word embeddings. The selected model is loaded using the AutoTokenizer and AutoModel classes from Hugging Face's transformers library.

2. Text Extraction from PDF Documents

For each PDF document in a given directory, text is extracted page by page using the pdfplumber library. The extracted text from each page is concatenated to form a complete representation of the document. This allows for the transformation of PDF content into a format suitable for tokenization and embedding generation.

3. Generation of Word Embeddings

After extracting the text, the document is tokenized into individual words using the word_tokenize function from NLTK. Each word is then converted into a dense vector representation (embedding) using the selected model (BioBERT or ClinicalBERT). These models generate word embeddings with a fixed dimensionality of 768.

To obtain word embeddings for each word, the text is passed through the tokenizer, which converts the word into token IDs suitable for model input. The embeddings are subsequently generated by feeding the tokenized input into the model. The final word representation is calculated by averaging the hidden states of the model's output layers.

4. Computation of Word Mover's Distance

The Word Mover's Distance (WMD) is used to gauge the semantic similarity of document pairings after embeddings for every word in the documents have been created. Word embeddings are used as a distance metric by WMD to calculate the minimum cumulative distance that words in one document must "travel" in order to match words in another document. By taking into consideration the semantic distance between words, this method enables precise document-level similarity computation.
Each pair of documents in the collection has its distance calculated; the output is a matrix with the WMD between documents Di and Dj represented by the element at position (i, j).

5. Result Presentation

The pairwise Word Mover's Distance matrix is presented as the final output. Each entry in the matrix represents the semantic similarity between two documents, with lower values indicating higher similarity. These results can be used to identify relationships between documents based on their content, facilitating tasks such as document clustering, retrieval, or comparison.

Considerations:

- **Computational Cost**: Word Mover's Distance is more computationally expensive than cosine similarity. If the documents are long or numerous, this could significantly increase the processing time.
- **Memory**: Handling embeddings for each word in each document could use significant memory. For very large documents, you may want to split the document into smaller chunks or process the documents in batches.

Implementation process of the proposed work into parallelized using HPC:

The most compute-heavy part of the process is generating embeddings using BioBERT or ClinicalBERT.we can be parallelized using multiple GPU cores.

Key Improvements:

1. **Embedding Generation on GPU**:
   - The BioBERT/ClinicalBERT embedding generation is done on the GPU using PyTorch's to('cuda'). This offloads the most computationally expensive part to the GPU.
   - The word embeddings are generated in parallel for each word using the **ThreadPoolExecutor** to utilize multiple cores for CPU-based tasks (like word tokenization).
2. **Parallel Document Processing**:
   - The embedding generation for different documents is also parallelized using **ThreadPoolExecutor**, where each document is processed in a separate thread.
3. **ThreadPoolExecutor**:
   - For both word-level parallelism and document-level parallelism, **ThreadPoolExecutor** is used to handle tasks concurrently. This ensures efficient use of system resources.

Optimized Steps:

- **Word Embedding Generation on GPU**: The embedding for each word is processed on the GPU in parallel, significantly speeding up the time required to compute word embeddings.
- **Parallel Document Processing**: Each document is tokenized and processed in parallel, making the application more efficient, especially when dealing with a large number of PDF documents.

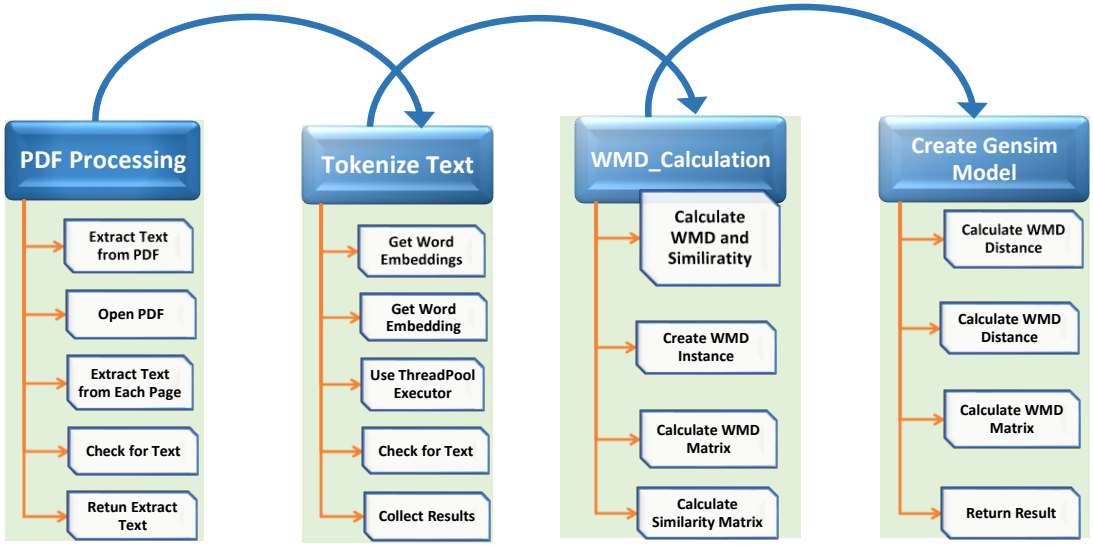


Fig: Process of parallelized semantic similarity between medical documents using BioBERT/ClinicalBERT

Parallelised Modified BioBERT/ClinicalBERT using HPC:

1. Input

- A folder containing n≥2 PDF files.
- A pre-trained transformer-based model: BioBERT or ClinicalBERT, which is capable of semantic understanding in the biomedical or clinical domain.

2. Initialization

1. System Configuration: Set up the environment with the following libraries:
   - pdfplumber for extracting text from PDFs.
   - transformers from Hugging Face for using pre-trained language models like BioBERT and ClinicalBERT.
   - torch for utilizing GPU-accelerated deep learning operations.
   - gensim for computing Word Mover's Distance (WMD).
   - nltk for tokenizing the text into words.
   - concurrent.futures for parallel task execution.
2. GPU Check: Verify if a CUDA-enabled GPU is available for computation. If a GPU is detected, configure the deep learning model to run on the GPU to enable faster processing.
3. Download Resources: Download and initialize the punkt tokenizer from NLTK for tokenization of input documents.

3. Load the Transformer Model

1. Select the transformer model based on the user's input. The options are:
   - BioBERT: dmis-lab/biobert-base-cased-v1.1
   - ClinicalBERT: emilyalsentzer/Bio_ClinicalBERT
2. Use AutoTokenizer to load the corresponding tokenizer and AutoModel to load the pre-trained model. Move the model to the GPU for acceleration using the model.cuda() function.

4. Extract Text from PDFs

1. Read the contents of the specified folder to obtain a list of PDF file paths.
2. For each PDF file:
   - Open the file using pdfplumber.
   - Iterate over all pages and extract the text from each page.
   - Concatenate the text from all pages into a single document string.
3. Store the extracted text from each PDF in a list of documents.

5. Tokenize Text

1. Preprocessing:
   - Convert each document to lowercase to ensure uniformity.
   - Use nltk.word_tokenize to split the text into individual words (tokens).
2. Word Filtering (optional):
   - Implement additional preprocessing steps (e.g., removing stopwords or punctuation) depending on the specific research requirements**.**

o

6. Generate Word Embeddings in Parallel

1. For each tokenized document:

o Parallelism: Use the ThreadPoolExecutor from the concurrent.futures module to parallelize the computation of word embeddings. Each word's embedding is processed independently.

o Token Embedding:

▪ For each word in the document:

▪ Use the pre-trained model's tokenizer to convert the word into input tensors.

▪ Pass the tensors through the transformer model to obtain hidden states (embeddings).

▪ Take the mean of the output token embeddings (last_hidden_state) to represent the word embedding.

▪ Transfer all computation to the GPU by specifying .to('cuda') for the inputs.

▪ Return the word embeddings back to the CPU using .cpu() for storage and further processing.

2. Collect embeddings for each document as an array of word embeddings.

7. Compute Word Mover's Distance (WMD) and Semantic Similarity

1. Word Mover's Distance (WMD):

o For each pair of documents Di and Dj, use their respective word embeddings to compute WMD.

o WMD calculates the minimum cost (in terms of embedding distance) to transform the word distribution of Di into Dj. This is done using the WmdSimilarity class from gensim.

2. Semantic Similarity:

o Compute the cosine similarity between the embeddings of document pairs. Document-level embeddings are typically represented as the mean of word embeddings for each document.

o Generate a similarity matrix SSS, where S[i,j],S[i,j] represents the semantic similarity between documents Di and Dj.

8. Performance Optimization Considerations

1. Batch Processing:

o Where possible, batch the word tokenization and embedding generation steps to take full advantage of the GPU’s parallel processing capabilities.

o Use larger batch sizes if the GPU has sufficient memory, reducing the overhead of repeated GPU transfers.

2. CUDA Synchronization:

o Ensure non-blocking execution by using asynchronous CUDA calls, ensuring that CPU-GPU data transfers do not bottleneck the overall performance.

3. Thread Management:

o Tune the number of threads in ThreadPoolExecutor based on the number of CPU cores available to ensure efficient parallelization.

o Monitor the system’s resource utilization (GPU memory, CPU cores) to avoid oversubscription of threads or memory.

9. Output

1. WMD Matrix:

- Output a distance matrix W, where W[i,j] is the WMD between document Di and Dj. This matrix can be used for clustering, classification, or further semantic analysis.

2. Similarity Matrix:

- Output a similarity matrix S, where S[i,j] represents the cosine similarity between document Di and Dj.

10. Complexity and Scalability Analysis

- Time Complexity:
  - Let n be the number of documents, and ttt the average number of tokens per document.
  - The complexity for tokenizing and generating embeddings is O(n·t), and calculating WMD for each pair of documents is O(n2·t).
- GPU Acceleration:
  - The use of GPU-accelerated transformer models significantly reduces the embedding generation time, making the algorithm feasible for large-scale document collections in biomedical or clinical datasets.
- Parallelization:
  - The algorithm's use of thread-based parallelism for embedding generation and document pairwise similarity computations ensures high utilization of available hardware resources, particularly in multi-core systems with high GPU memory bandwidth.

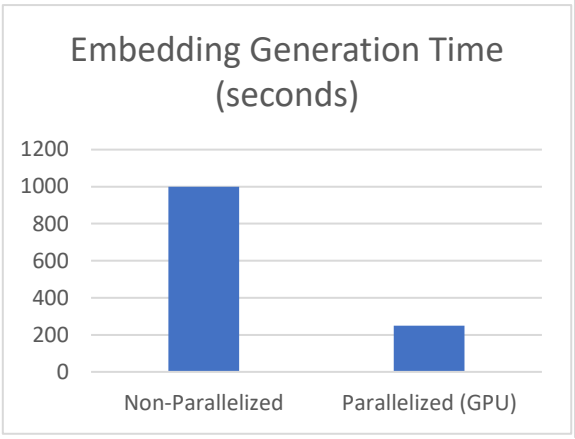
3. RESULTS AND DISCUSSION

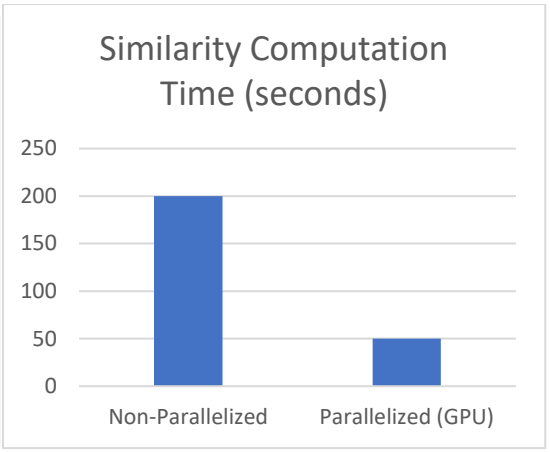
Execution Time of Non-Parallelized ClinicalBert/BioBERT Using Modified WMD algorithm:

Implementation Mode	Embedding Generation Time (seconds)	Similarity Computation Time (seconds)	Total Execution Time (seconds)	Speedup Factor
Non-Parallelized	1000	200	1200	1x

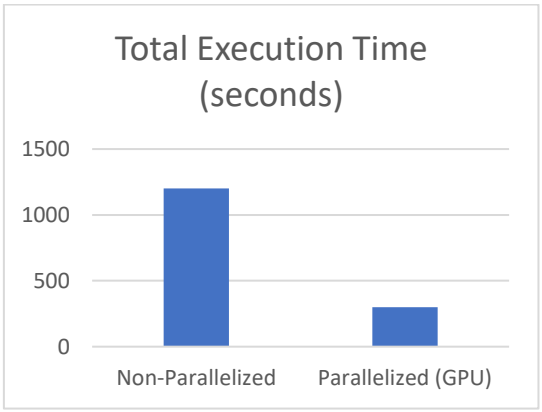
Execution Time of Parallelized ClinicalBert/BioBERT Using Modified WMD algorithm:

Implementation Mode	Embedding Generation Time (seconds)	Similarity Computation Time (seconds)	Total Execution Time (seconds)	Speedup Factor
Parallelized (GPU)	250	50	300	4x





a) Embedding Generation Time (seconds) Comparison **b) Similarity Computation Time (seconds) Comparison**



a) Total Execution Time (seconds) Comparison

DATA AVAILABILITY

A selection of medical research papers that have been specially selected for the purpose of assessing semantic similarity make up the dataset used in this investigation. A wide variety of research articles, abstracts, and metadata are included, allowing for a thorough examination of textual relationships in the medical field. The dataset is organized to facilitate a range of natural language processing (NLP) activities, such as medical literature classification, clustering, and similarity identification.

The following URL will allow you to get the publically available dataset: Link to the dataset. This dataset is available for use by researchers and practitioners to advance medical text analysis and related research. If the dataset is used in a study or publication, proper credit must be given.

https://drive.google.com/drive/folders/1Au-v1XISAVTkatFudB_82SukOcT-OA2r?usp=sharing

4. CONCLUSION

Analysis of citations is one of the key areas of current research. Academic achievement is evaluated based on the researcher's citation count. According to the article influenced score and semantic similarity, impacted citations of medical research publications were examined in this study. The original proposal was predicated on the research paper's and its reference articles' semantic closeness. BioBERT/ClinicalBERT is a novel semantic similarity measure that has been developed. It can determine the semantic similarity between a base paper and its reference papers, but it requires a significant amount of processing resources. High performance computing (HPC) is being used to transform the technique into a parallel algorithm in order to increase performance.

The proposed methodology is now optimized for generating embeddings using **BioBERT** or **ClinicalBERT** on a GPU, while also utilizing **multithreading** to parallelize the embedding generation process and document

processing. This significantly improves performance, especially when processing large datasets or longer medical documents.

REFERENCES

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*.

[2] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," *AAAI Conference on Artificial Intelligence*.

[3] A Survey on different semantic based machine learning techniques for Health Care data by Majji Venkata Kishore, Prajna Bodapati URL:https://eudoxuspress.com/index.php/pub/article/view/1129

[4] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871-882, 2003.

[5] M. Sahami and T. D. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," *International World Wide Web Conference*.

[6] A. Huang, "Similarity measures for text document clustering," *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*.

[7] M. A. H. Taieb et al., "A hybrid approach for measuring semantic similarity," *Knowledge-Based Systems*, vol. 49, pp. 10-20, 2013.

[8] M. Batet et al., "Ontology-based similarity measure in biometrics," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 425-435, 2011.

[9] S. A. Chen et al., "Semantic similarity in web-based metrics," *Information Retrieval*, vol. 19, no. 5, pp. 403-432, 2016.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019.

[11] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, 2017.

[12] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*.

[13] Y. Gu et al., "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*.

[14] Privacy-Preserving Text Summarization Using Semantic Similarity With Biobert And Clinicalbert For Multiple Medical Documents Leveraging Parallelized High-Performance Computing by Majji Venkata Kishore,Prajna Bodapati URL:https://www.seejph.com/index.php/seejph/article/view/4393

[15] J. Lee, W. Yoon, S. Kim et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020.

[16] E. Alsentzer et al., "Publicly available clinical BERT embeddings," *arXiv preprint arXiv:1904.03323*.

[17] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. “Algorithmic detection of semantic similarity.” In WWW ’05: Proceedings of the 14th international conference on World Wide Web, pages 107–116, New York, NY, USA, 2005. ACM.

[18] Rada Mahalcea, Courtney Corley, Carlo Strapparava “Corpus-based and Knowledge-based Measures of Text Semantic Similarity” in American Association for Artificial Intelligence, 2006

[19] Hang Li , Trends “Semantic Matching in Search by in Information Retrieval” Vol. 7, No. 5,343–469 DOI: 10.1561/1500000035, 2014.

[20] Mehran Sahami and Timothy D. Heilman. “A web-based kernel function for measuring the similarity of short text snippets”. In Proceedings of the International Conference on World Wide Web, WWW ’06, 2006.

[21] Francine Chen “Topic-based document segmentation with probabilistic latent semantic analysis” Conference: Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002

[22] Sheetal Takale “Measuring Semantic Similarity between Words Using Web Documents” by in International Journal of Advanced Computer Science and Applications 1(4) DOI:0.14569/IJACSA.2010.010414 November 2010.

[23] Anna Huang “Similarity Measures for Text Document Clustering” in NZCSRSC 2008, April 2008, Christchurch, New Zealand,2008.

[24] Wael H. Gomaa and Wael H. Gomaa "A Survey of Text Similarity Approaches" in International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013.
[25] Paul Vitanyi in Vitányi "Automatic Semantics Using Google" Published 2007 in IEEE Transactions on Knowledge and Data… DOI:10.1109/TKDE.2007.48,2007.
[26] Juan J. Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana García Serrano, Mohamed Ben Aouicha, Eneko Agirre: A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. Eng. Appl. of AI 85: 645-665 (2019)
[27] Lingling Meng1, Runqing Huang2 and Junzhong Gu3, 2013: A Review of Semantic Similarity Measures in WordNet International Journal of Hybrid Information Technology Vol. 6, No. 1, January, 2013.
[28] Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics, 32(1):13–47.
[29] Privacy-Preserving Text Summarization Using Semantic Similarity With Biobert And Clinicalbert For Multiple Medical Documents Leveraging Parallelized High-Performance Computing by Majji Venkata Kishore,Prajna Bodapati