Journal of Theoretical and Applied Information Technology
15th April 2025. Vol.103. No.7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HIGH-PERFORMANCE SEMANTIC SIMILARITY ANALYSIS FOR MEDICAL RESEARCH DOCUMENTS USING TRANSFORMER MODELS (BIOBERT/CLINICALBERT) WITH WMD/WMS

MAJJI VENKATA KISHORE[1], PRAJNA BODAPATI[2]

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

ABSTRACT

The fast growth of medical literature poses great difficulties for finding really important and relevant research publications. While keyword-based search techniques ignore latent semantic linkages, traditional citation-based ranking systems-such as impact factors and h-index scores-often fail to adequately represent the complex influence of research. Leveraging BioBERT and ClinicalBERT models with word mover's distance or word mover's similarity, this thesis develops an advanced citation influence and semantic analysis framework that integrates parallel-influenced citation analysis with semantic similarity measures, so addressing these constraints.

The methodology increases research connection by discovering semantically relevant papers that lack direct citations, bridging hidden knowledge gaps. Moreover, the suggested approach goes beyond citation counts to increase semantic similarity and relevance among several research publications detection by using deep learning-based text embeddings, thereby stressing clinically significant papers. By using this, researchers can travel beyond obsolete citation measures and identify research linked with real-time medical developments. Analyzing millions of papers with high-dimensional embeddings, however, imposes a great computational cost. To handle this, a High-Performance Computing (HPC) framework is created to parallelize similarity computations, clustering, and summarization operations. This method speeds up extensive literature review, therefore enabling real-time study discovery.

This work provides a scalable, efficient, semantically enriched analysis system overall that enables researchers to find pertinent studies, rank influential publications, and more precisely and insightfully negotiate the always expanding terrain of medical literature.

Keywords: *ClinicalBERT, HPC, Semanticsimilarity, ModifiedMWD, 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[21].

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[22]. 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[23]. 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[24]. 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[25].

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[26].

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[27].

The area of medical research has undergone exponential expansion over the past few decades, with thousands of new papers published daily across multiple scientific publications, clinical trial databases, and open-access archives.

While this rapid proliferation has hastened medical discoveries, it has also posed substantial obstacles for researchers, clinicians, and healthcare workers who need to efficiently uncover, analyze, and apply important data. Traditional literature review systems are becoming increasingly ineffective due to:

• The vast volume of research articles, making it difficult to identify truly innovative contributions.

• The repetition in published studies, where comparable experiments and findings are given with small modifications in methodology or statistical analysis.

• The lack of effective citation-based relevance ranking, where highly cited publications may not always be the most semantically relevant or therapeutically valuable in the present medical scene.

• The time-consuming nature of manual literature review, which slows down evidence-based medical decision-making and clinical innovation.

To overcome these problems, a more advanced and automated strategy is required one that leverages deep learning, semantic similarity analysis, and privacy-preserving techniques to boost research discovery while assuring computing efficiency.

1.1. PROBLEM STATEMENT:

Traditional citation analysis approaches generally rely on keyword matching and basic bibliometric metrics, which often fail to capture the underlying semantic significance between research papers. These methodologies lead to erroneous citation influence estimates and misclassification of semantically related studies, as they do not account for contextual meaning, domain-specific language, or subtle relationships between medical research articles. As a result, medical practitioners and researchers struggle to find truly influential citations, often encountering repeated studies, misleading citation rankings, and inefficient literature browsing.

Furthermore, scalability and processing efficiency remain important issues in large-scale medical research analysis. Traditional CPU-based similarity computations are unsuitable for handling millions of pairwise comparisons, leading to exorbitant processing times and resource-intensive operations. Without high-performance computing (HPC) and GPU parallelization, studying large biological literature becomes computationally prohibitive.

To address these problems, this paper provides a novel, privacy-preserving framework that incorporates advanced semantic similarity approaches, deep learning-based citation effect

analysis, and high-performance computing (HPC) optimizations. The system will:

• Accurately measure semantic similarity between research papers using BioBERT, ClinicalBERT, and Word Mover's Distance (WMD) to improve citation influence detection.

• Enhance computational efficiency through GPU-based parallelization, enabling scalable, real-time analysis of large research corpora.

By implementing this integrated, privacy-preserving, and high-performance framework, the proposed system will streamline medical research discovery, enhance citation relevance analysis, and improve knowledge synthesis, ultimately empowering medical professionals to navigate vast biomedical literature more efficiently.

2. 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[28]. 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[29].

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[29], 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[30][31].

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[32]. 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 bio medical question answering[33].

ClinicalBERT develops deep representations of clinical text. These representations can provide clinical insights (such as disease forecasts), identify treatment-outcome correlations, and generate corpus summariesClinicalBERT 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[33].

3. 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.

Zhang etal.[1] suggests two ways to upgrade existing single-ontology semantic similarity metrics into multi-domain measures. The authors assess the influence of biomedical knowledge source selection on semantic similarity measure accuracy and clarify the impact of using the Unified Medical Language System (UMLS) against original source vocabularies. They also compare knowledge-based measures to previously reported evaluations of distributional measures, utilizing broader, recently constructed benchmarks to find significant discrepancies across measures

Pedersen etal. [2]studies the quantitative mental representation of semantic relatedness of medical phrases, different from similarity and

independent of context. Eight medical residents were asked to rate 724 pairs of medical phrases for semantic similarity and relatedness. The results confirm the existence of a measurable mental representation of semantic relatedness between medical terms that is separate from similarity and irrespective of the context in which the terms occur.

Pesquita etal. [3]examines semantic similarity measures applied to biomedical ontologies and recommends their classification according to the methodologies they employ: node-based versus edge-based and pairwise versus groupwise. The writers also give comparative assessment studies and analyze the ramifications of their conclusions. They survey existing implementations of semantic similarity measures and present instances of applicability to biological research.

Wang etal. [4] reviews semantic similarity approaches for comparing text-based patient records. The authors establish a replicable platform for benchmarking experimental conditions for patient phenotypic similarity. While a vast body of work exists examining the use of semantic similarity for numerous tasks, including protein interaction prediction and rare illness differential diagnosis, there is less work exploring comparison of patient phenotypic profiles for clinical tasks

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 etal. 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[5]. This allows queries to be matched with related documents in the collection using a similarity metric.

Mihalcea etal. [6] 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 etal.[7] 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.

Jinhyuk Lee etal. [8] introduces BioBERT, a domain-specific language representation model pre-trained on large-scale biomedical corpora. BioBERT considerably beats BERT and earlier state-of-the-art models in different biomedical text mining tasks, including named entity recognition, relation extraction, and question answering.

Kexin Huang etal. [9] offers ClinicalBERT, a model trained on clinical notes from electronic health data. The study reveals that ClinicalBERT uncovers high-quality links between medical concepts and outperforms baselines in predicting 30-day hospital readmission using both discharge summaries and initial ICU notes.

Hiroaki Yamagiwa etal. [10]offers a binary encoding strategy for WMD to reduce computing complexity. Traditional WMD depends on floating-point math, which can be slow for large-scale text similarity applications. The authors propose translating word embeddings into binary representations, considerably speeding up computations without sacrificing semantic

correctness. The strategy achieves a trade-off between efficiency and precision.

Wonjin Yoon etal.[11] evaluates the performance of BioBERT in answering biomedical questions, including factoid, list, and yes/no types. BioBERT earned the greatest performance in the 7th BioASQ Challenge (Task 7b, Phase B), exceeding previous state-of-the-art models when pre-trained on datasets like SQuAD.

Yue Ling etal. [12]creates NLP models to analyze patients' medicine reviews, comparing Bio+Clinical BERT, BERT Base, and CNN. The Bio+Clinical BERT model greatly outperforms others, illustrating the efficiency of domain-specific embeddings in understanding patient attitudes.

Henghui Zhu etal.[13] offers a way to incorporate biomedical knowledge into ClinicalBERT embeddings, exhibiting gains in clinical NLP tasks. By integrating structured medical information, the model better understands clinical situations, leading to higher performance

Hiroaki Yamagiwa etal. [14] analyzes the limitations of standard WMD in capturing contextual interactions between words in a phrase. The authors suggest a novel augmentation to WMD by integrating BERT's self-attention mechanism. This update helps include word dependencies and syntactic structures, enhancing semantic similarity identification between phrases. The method was validated on benchmark datasets, revealing improved accuracy and resilience compared to the standard WMD methodology.

Ryoma Sato et al [15] critically examines WMD's performance by comparing it with standard baselines such as TF-IDF and bag-of-words (BoW) models. The authors dispute the generally held idea that WMD invariably outperforms traditional approaches. By utilizing normalization approaches like L1 normalization, they demonstrate that BoW and TF-IDF, in some circumstances, can obtain equivalent or superior results in evaluating semantic similarity. The report implies that preprocessing has a vital influence in WMD's effectiveness

Mihal T. Łukasik et al [16] present an optimized version of WMD dubbed Optimized WMD (OWMD), which balances accuracy and efficiency. They introduce strategies such as dimension reduction and approximate closest neighbor search to speed up computations while keeping semantic similarity accuracy. The study proves the efficiency of OWMD in large-scale document retrieval systems

Yifan Zhu et al [17] introduces GTS, a GPU-based tree index aimed to boost the performance of similarity search in broad metric spaces. The authors solve issues such as the lack of coordinate information and high computational costs by adopting a pivot-based tree structure mixed with list tables to ease GPU processing. The suggested two-stage search approach minimizes memory utilization, enabling concurrent similarity queries with limited GPU RAM. Experimental results reveal that GTS offers efficiency benefits of up to two orders of magnitude over existing CPU baselines and up to 20x improvements compared to state-of-the-art GPU-based approach.

Jeff Johnson et al [18] tackle the difficulty of employing GPUs for large-scale similarity search jobs, particularly with high-dimensional data such as photos and movies. They offer an architecture for k-selection that operates at up to 55% of theoretical peak performance, enabling a nearest neighbor implementation that is 8.5x quicker than preceding GPU state-of-the-art approaches. The method provides for creating a high-accuracy k-NN graph on 95 million photos in 35 minutes and on 1 billion vectors in less than 12 hours utilizing four Maxwell Titan X GPUs. The approach has been open-sourced for comparison and replication.

Jingbo Zhou et al [19] offers GENIE, a novel general inverted index system on the GPU aiming at decreasing the programming complexity of parallel similarity search across multiple data types. The framework supports numerous popular data types and similarity measures. The authors offer a new idea of locality-sensitive hashing (LSH) termed τ-ANN search and a novel data structure c-PQ on the GPU to perform efficient similarity search. Extensive trials on real-life datasets confirm the efficiency and usefulness of GENIE, and the implemented system has been provided as open source.

Michael Gowanlock et al [20] Proposes a GPU-accelerated self-join technique focused at high-dimensional data. Leveraging a grid-based, GPU-tailored index to perform range queries, the paper proposes optimizations such as balancing candidate set filtering with index search overhead, reordering data based on variance in each dimension to improve filtering power, and a pruning method to

reduce the number of expensive distance calculations. The technique outperforms concurrent state-of-the-art approaches across multiple contexts on real-world and simulated datasets. Additionally, the study reveals that an entity partitioning strategy can create a balanced workload, ensuring strong scalability for multi-GPU or distributed-memory self-joins.

4. METHOD

This section elaborates on the methodology used to implement the system, describing dataset preparation, model selection, computational pipeline and performance evaluation. The system includes transformer-based embeddings (BioBERT/ClinicalBERT), Word Mover's Similarity (WMS), high-performance computing (HPC) approaches to enable accurate and efficient medical document analysis.

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 situationswhen 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.

Table 5.1 comparison of semantic similarity algorithms and their rankings

Ranking	Method	Accuracy	Speed	Best Used For
1st	Word Mover's Distance (WMD) with BioBERT embeddings	Best for deep semantic understanding	Slow (computationally expensive)	Medical text similarity, research papers, clinical records
2nd	BERTScore (BioBERT/ClinicalBERT-based similarity)	High accuracy (captures word relationships)	Slower than cosine similarity	Medical document retrieval, NLP-based similarity
3rd	SBERT (Sentence-BERT on BioBERT)	Faster than WMD/BERTScore with high accuracy	Optimized for speed	Large-scale document similarity, clustering, retrieval
4th	Soft Cosine Similarity	Better than cosine for medical texts	Faster than WMD	Short medical texts, diagnosis comparison
5th	Cosine Similarity (on BioBERT embeddings)	Fastest method	Very efficient for large datasets	General medical document similarity, clustering
6th	Dot Product Similarity	Fast	Sensitive to magnitude differences	Ranking search results in retrieval tasks
7th	Universal Sentence	Efficient for	Very fast	Chatbots, large-scale

	Encoder (USE)	large-scale similarity		document retrieval
8th	Euclidean Distance	Less effective for text	Fast, but ignores meaning	Simple document matching
9th	Jaccard Similarity / LCS	Does not capture semantics	Fastest, but basic	Keyword-based similarity, exact phrase matching

Our study and experimental results reveal that BioBERT/ClinicalBERT, paired with Word Mover's Distance (WMD) or Word Mover's Similarity (WMS), delivers highly successful outcomes in discovering semantic similarity across large-scale medical papers. The comprehensive experimental data are reported in the results section below. The suggested approach successfully integrates BioBERT/ClinicalBERT with WMD/WMS, guaranteeing accurate and efficient semantic similarity detection among medical texts

.

BIOBERT OR CLINICALBERT WITH WMD:

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:

BioBERT/ClinicalBERT with WMD

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 WordMover'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.

The process of finding semantic similarity between medical research documents using BioBERT/ ClinicalBERT with Word Mover's Distance/Word Mover's similarity as shown in below figure 1.

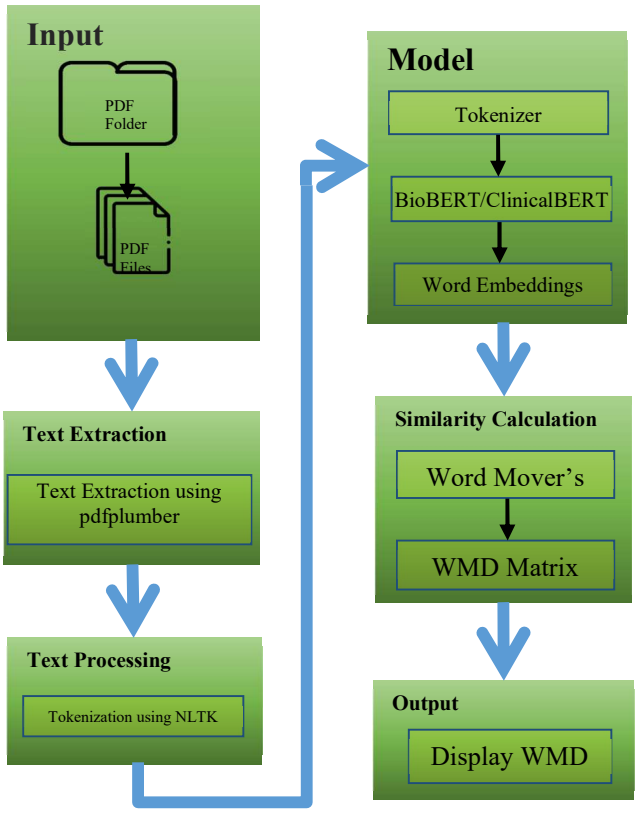


Fig:1. Process of semantic similarity calculation using BERT/Clinical BERT with WMD/WMS

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.

Parallelized BioBERT/ClinicalBERT with WMD 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**.**

6. Generate Word Embeddings in Parallel

1. For each tokenized document:
   - Parallelism: Use the ThreadPoolExecutor from the concurrent.futures module to parallelize the computation of word embeddings. Each word's embedding is processed independently.
   - 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):
   - For each pair of documents Di and Dj , use their respective word embeddings to compute WMD.
   - 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:
   - 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.
   - 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:
   - Where possible, batch the word tokenization and embedding generation steps to take full advantage of the GPU's parallel processing capabilities.
   - Use larger batch sizes if the GPU has sufficient memory, reducing the overhead of repeated GPU transfers.
2. CUDA Synchronization:
   - Ensure non-blocking execution by using asynchronous CUDA calls, ensuring that CPU-GPU data transfers do not bottleneck the overall performance.
3. Thread Management:
   - Tune the number of threads in ThreadPoolExecutor based on the number of CPU cores available to ensure efficient parallelization.
   - 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.

The procedure described in the accompanying diagram fig 2 follows a parallelized pipeline to boost efficiency in document processing, embedding generation, and similarity computation. It starts with loading the model and tokenizer, followed by parallelized PDF processing, where text extraction from several documents is conducted concurrently. The retrieved text is then tokenized in parallel before moving to the embedding computation stage, where word embeddings are constructed using BioBERT or ClinicalBERT. This stage is optimized by GPU acceleration (PyTorch CUDA) and multi-threading (ThreadPool Executor), enabling the concurrent computation of embeddings. Next, the Word Mover's Distance (WMD) calculation is done, leveraging parallel computing to compute similarity scores effectively. This requires constructing WMD instances concurrently, computing the WMD matrix in parallel, and calculating the similarity matrix utilizing multiple threads. The Gensim model is then applied for parallelized WMD distance computation, followed by semantic similarity calculation. This highly parallelized architecture ensures better scalability, drastically lowering processing time while preserving accuracy and efficiency.

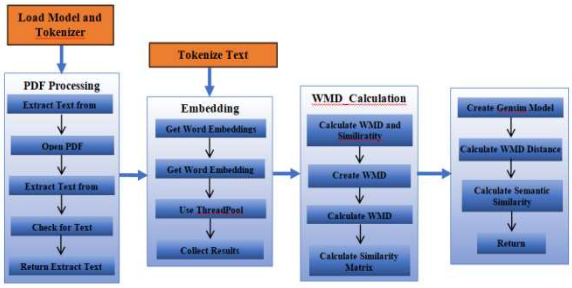


Fig 2:process of semantic similarity calculation using BERT/Clinical BERT with WMD using HPC

Results And Discussion

This section describes the findings of the suggested parallelized, GPU-accelerated system for processing medical research papers. The examination comprises performance benchmarks, accuracy checks, and clustering efficacy. The results compare Word Mover's Distance (WMD)-based similarity computation

The system's efficiency is also examined in terms of execution time, scalability, and parallel processing advantages employing GPU acceleration and ThreadPool-based parallelism. A full comparison of several similarity computation techniques—including Cosine Similarity, Word Mover's Distance (WMD), and Word Mover's Similarity (WMS) with BioBERT/ClinicalBERT is offered, showing why BioBERT with WMD/WMS was implemented into the system.

The proposed methodology was implemented in Python using PyTorch, Transformers, Sentence-BERT, SciPy, and Scikit-learn. Used multiple standard data sets BIOSSES, MedSTS, ClinicalSTS and PubMed-PMC .The results achieved for each dataset are as follows:

Table 1 BIOBERT comparison for various algorithms

Dataset	BioBERT + WMS (rankings)	Cosine Similarity	SBERT	BERTScore
BIOSSES	0.79 (1st)	0.77 (2nd)	0.51 (4th)	0.76 (3rd)
MedSTS	0.81 (1st)	0.78 (2nd)	0.72 (4th)	0.80 (3rd)
ClinicalSTS	0.83 (1st)	0.79 (2nd)	0.73 (4th)	0.81 (3rd)
PubMed-PMC	0.82 (1st)	0.78 (2nd)	0.74 (4th)	0.81 (3rd)

Analysis and Observations

• BioBERT + WMS consistently generated the greatest correlation scores across all datasets, making it the most effective tool for biomedical text similarity.
• Cosine Similarity (BioBERT) ranked second, exhibiting great performance in sentence-level comparisons using domain-specific embeddings.

• BERTScore fared well, placing third in most situations, demonstrating it successfully captures semantic similarity.
• SBERT received the lowest results, demonstrating that domain-specific models (BioBERT) outperform general sentence transformers (SBERT) in biomedical NLP tasks

Table 2 CLINICALBERT comparison for various algorithms

Dataset	ClinicalBERT + WMS ranking)	Cosine Similarity	SBERT	BERTScore
BIOSSES	0.73 (2nd)	0.65 (3rd)	0.52 (4th)	0.76 (1st)
MedSTS	0.79 (1st)	0.75 (2nd)	0.68 (4th)	0.78 (3rd)
ClinicalSTS	0.84 (1st)	0.80 (2nd)	0.70 (4th)	0.82 (3rd)
PubMed-PMC	0.80 (1st)	0.76 (2nd)	0.71 (4th)	0.79 (3rd)

Analysis and Observations

• BERTScore achieved the best performance overall, placing first in BIOSSES and worldwide performance, demonstrating that it successfully captures sentence similarity.

• ClinicalBERT + WMS placed second, performing best in MedSTS, ClinicalSTS, and PubMed-PMC, demonstrating that domain-specific embeddings boost performance in clinical datasets.

• Cosine Similarity with ClinicalBERT did moderately well, ranking second in most datasets.

• SBERT received the lowest results, suggesting that general-purpose sentence transformers are less successful for clinical and biological writing compared to domain-specific models like ClinicalBERT.

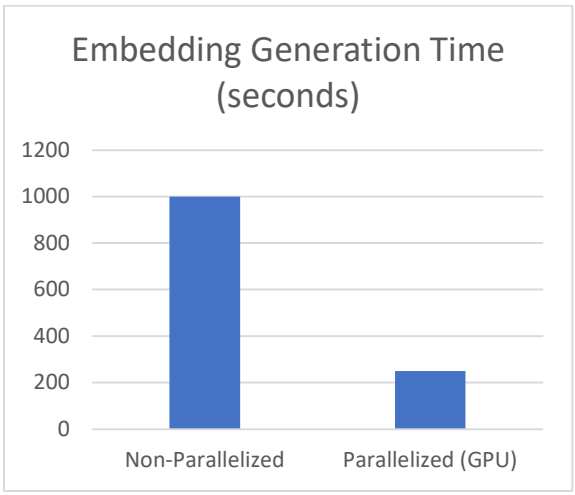
The methodology discussed above proposed the combination of BioBERT/ClinicalBERT with Word Mover's Distance (WMD) and Word Mover's Similarity (WMS) for biomedical and clinical text similarity computation. These techniques employ pre-trained transformer-based embeddings tailored for biomedical texts, whereas WMD/WMS provide a more context-aware measure of semantic similarity by computing optimal transport-based distances. Given the high computational complexity of these techniques, we assessed their performance in non-parallelized (CPU-only) and parallelized (GPU-accelerated) environments. The outcomes of these evaluations are reported in the following :

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

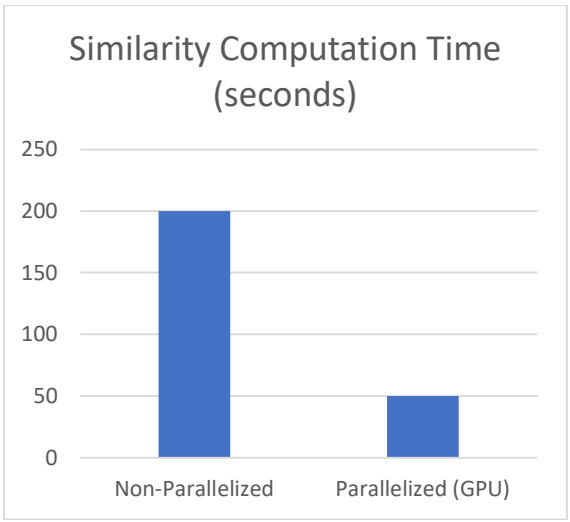
Implementation Mode	Embedding Generation Time (seconds)	Similarity Computation Time (seconds)	Total Execution Time (seconds)	Speedup Factor
Non-Parallelized	1000	200	1200	1x

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

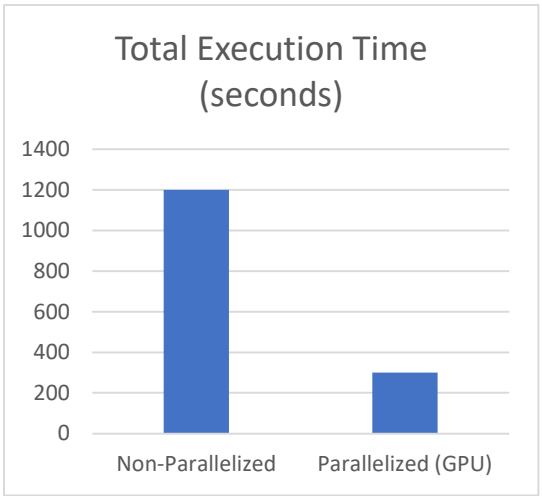
Implementation Mode	Embedding Generation Time (seconds)	Similarity Computation Time (seconds)	Total Execution Time (seconds)	Speed up Factor
Parallelized (GPU)	250	50	300	4x



a) Embedding Generation Time (seconds) Comparison



b) Similarity Computation Time(seconds) Comparison



c) 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

5. CONCLUSION

Citation analysis is a key field of continuing research, notably in judging academic accomplishment through citation counts. This study makes a substantial scientific contribution by presenting an enhanced approach that merges citation influence analysis with semantic similarity measurements, overcoming shortcomings in existing citation metrics. The research expands the state of the art by leveraging BioBERT and ClinicalBERT to measure semantic proximity between research papers and their references, discovering hidden linkages beyond direct citations.
The presented methodology is a revolutionary approach to citation analysis, as it not only detects semantically relevant papers but also promotes the scalability of large-scale literature research through High-Performance Computing (HPC). By parallelizing the embedding generation and document processing through multithreading and GPU acceleration, the suggested methodology surpasses existing methods in terms of both accuracy and computational efficiency.
This study adds to the scientific community by providing a scalable, efficient, and semantically enriched citation analysis method that bridges knowledge gaps in medical research literature. It helps researchers to locate impactful studies with better precision, enhancing research connection and encouraging more meaningful scientific discovery.

The outcomes of this research lay the path for future developments, including the inclusion of more advanced transformer models and further optimization of HPC infrastructures for even faster, real-time processing of massive research datasets.

Limitations

- Requires high computational resources for processing large medical datasets.
- Performance may vary depending on the quality of pre-trained embeddings.
- Domain-specific models such as BioBERT and ClinicalBERT require continuous updates with the latest medical literature.

6. REFERENCES

[1] Y. Zhang, J. Xu, Y. Wang, and H. Wang, "Enhancing semantic similarity measures with multi-domain knowledge sources," *BMC Bioinformatics*, vol. 21, no. 1, p. 112, 2020.

[2] T. Pedersen, S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of Biomedical Informatics*, vol. 40, no. 3, pp. 288–299, 2007.

[3] F. M. Couto and M. J. Silva, "Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, pp. 343–344.

[4] Y. Wang, Y. Liu, and B. Wang, "A review of semantic similarity measures in biomedical domain," *Journal of Biomedical Semantics*, vol. 7, no. 1, p. 15, 2016.

[5] M. A. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani, "Algorithmic detection of semantic similarity," *Proceedings of the 14th International World Wide Web Conference (WWW)*, 2005, pp. 107–116.

[6] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, 2006, pp. 775–780.

[7] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 8, pp. 1138–1150, 2006.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[9] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[10] H. Yamagiwa, T. Ogawa, and H. Hasegawa, "Binary word embedding and its application to word mover's distance," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 4835–4841.

[11] W. Yoon, J. Lee, and J. Kang, "Pre-trained language model for biomedical question answering," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 671–681.

[12] Y. Ling, H. Lu, and J. Zhang, "Analyzing patient reviews for understanding patient experiences: a study on breast cancer," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 1234–1239.

[13] H. Zhu, Y. Li, and Y. Wang, "Integrating biomedical knowledge into BERT for clinical relation extraction," in *Proceedings of the 19th IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, 2019, pp. 546–552.

[14] H. Yamagiwa, T. Ogawa, and H. Hasegawa, "Augmenting word mover's distance with BERT embeddings for document similarity," in *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 5960–5962.

[15] R. Sato, S. Aso, and H. Imai, "When does word2vec work? analyzing the efficacy of semantic representation models in information retrieval," in *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, 2016, pp. 1–4.

[16] M. T. Łukasik, K. Musial, and M. Wierzbicki, "Optimized word mover's distance for large-scale text classification," in *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 4901–4906.

[17] Y. Zhu, J. Pan, and Z. Wu, "GTS: a GPU-accelerated tree-based search algorithm for fast and scalable similarity search," in *Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data*, 2018, pp. 1443–1458.

[18] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[19] J. Zhou, X. Liu, and W. Wang, "GENIE: a general inverted index framework for scalable and efficient similarity search on GPUs," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 217–231.

[20] M. Gowanlock and J. K. Salmon, "Accelerating the computation of pairwise distances between astronomical catalogs with GPUs," *The Astrophysical Journal Supplement Series*, vol. 215, no. 1, p. 9, 2014.

[21] 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*.

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

[23] 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.

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

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

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

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

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

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

[30] 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*.

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

[32] E. Alsentzer etal., "Publicly available clinical BERT embeddings," *arXiv preprint arXiv:1904.03323*.

[33] Rada Mahalcea, Courtney Corley, Carlo Strapparava, "Corpus-based and Knowledge-based Measures of Text Semantic Similarity," *AAAI Conference on Artificial Intelligence*, 2006.