

Comparative analysis of deep Siamese models for medical reports text similarity

Dian Kurniasari^{1,2}, Mustofa Usman², Warsono², Favorisen Rosyking Lumbanraja³

¹Doctoral Program, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia

²Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia

³Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia

Article Info

Article history:

Received Feb 27, 2024

Revised Jul 25, 2024

Accepted Aug 6, 2024

Keywords:

Biomedical natural language processing
BioWordVec
Hybrid LSTM-CNN
Medical report
Semantic text similarity
Siamese Manhattan

ABSTRACT

Even though medical reports have been digitized, they are generally text data and have not been used optimally. Extracting information from these reports is challenging due to their high volume and unstructured nature. Analyzing the extraction of relevant and high-quality information can be achieved by measuring semantic textual similarity (STS). Consequently, the primary aim of this study is to develop and evaluate the performance of four models: Siamese Manhattan convolution neural network (CNN), Siamese Manhattan long short-term memory (LSTM), Siamese Manhattan hybrid CNN-LSTM, and Siamese Manhattan hybrid LSTM-CNN, in determining STS between sentence pairs in medical reports. Performance comparisons were conducted using Cosine Similarity and word mover's distance (WMD) methods. The results indicate that the Siamese Manhattan hybrid LSTM-CNN model outperforms the other models, with a similarity score of 1 for each sentence pair, signifying identical semantic meaning.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mustofa Usman

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung

Prof. Sumantri Brojonegoro street No.1, Gedong Meneng, Bandar Lampung, Lampung, Indonesia

Email: usman_alpha@yahoo.com

1. INTRODUCTION

Over the past decade, there has been abundant textual data across various fields, including biomedical research [1]. However, these valuable resources have often gone underutilized. For instance, much medical report text data remains archived and unexplored, even after digitization. This data's sheer volume and unstructured nature make extracting relevant, high-quality information difficult.

One practical approach to extracting relevant, high-quality information from medical report texts involves measuring semantic textual similarity (STS). STS is a foundational natural language processing (NLP) task that assesses how closely two sentences convey the same meaning [2]. In biomedical NLP (BioNLP), STS plays a pivotal role in ensuring the accurate interpretation and retrieval of information from biomedical documents [3]. That is due to its numerous direct applications in information extraction, such as in biomedical sentence search and classification, as well as indirect applications like biomedical question answering and document labelling [4]. The significance of STS in these applications is heightened by the fact that many biomedical terms can have different meanings depending on the text's context [5].

Researchers typically employ three main approaches to calculate STS: Corpus-based, knowledge-based, and string-based methods. Research by Sunilkumar and Shaji [6] indicates that the Corpus-based approach, in particular, is widely adopted and shows promising results. The Corpus-based method comprises two statistically distinct approaches. The first approach utilizes traditional statistical analysis techniques like latent semantic analysis (LSA), which counts word frequencies in the text but does not delve deeply into

semantic information extraction. On the other hand, the second approach employs deep learning (DL) to generate word embeddings that capture contextual meanings in text, assessing similarity using multilayer perceptrons (MLPs) to determine label similarities [7].

Well-known word embedding models like Word2Vec, GloVe, and FastText are commonly employed for STS and trained on general-domain corpora. However, applying Corpus-based STS methods in BioNLP research necessitates specialized approaches, such as using domain-specific biomedical corpora or biomedical knowledge sources. Despite these efforts, their adoption within the biomedical domain remains limited [8]. Furthermore, the annotation process in biomedical research demands the expertise of medical professionals [7], [8], leading to a scarcity of labelled datasets and hindering progress in biomedical STS.

Measuring the Cosine similarity (CS) and word mover's distance (WMD) between two sentences is widely recognized as the simplest and most frequently applied method for evaluating STS in unlabeled data. However, CS often exhibits lower performance levels [9]–[12]. Conversely, WMD methods struggle to differentiate sentences that share identical terms but possess distinct semantic meanings, primarily due to their disregard for word order [13].

Another widely adopted and highly regarded approach in this domain is the DL model utilizing the Siamese neural network architecture introduced by Mueller and Thyagarajan [14]. The Siamese model architecture involves two identical neural networks operating in parallel, extracting word representations from input vectors. The final output is typically compared using Cosine distance to determine STS [15], [16]. Subsequent research by Shi *et al.* [17] indicated that the Manhattan similarity metric offers faster convergence and higher accuracy than other metrics, including Cosine distance. Henceforth, this model is referred to as the Siamese Manhattan.

The issue tackled in this paper is the underutilization and difficulty of transforming large volumes of unstructured medical report text data into actionable information. The proposed solution in this paper is to implement the Siamese Manhattan architecture in DL to assess Biomedical STS using unlabeled leukemia medical report data and a Corpus-based strategy with biomedical domain-specific corpora known as BioWordVec. The model incorporates two distinct labelling processes utilizing the CS and WMD methods. The paper aims to demonstrate the effectiveness of the Siamese Manhattan model in accurately interpreting and retrieving information from biomedical documents, addressing the problem of underutilized textual data in the biomedical field.

The structure of this paper is as follows: section 2 will review related work on STS. Section 3 will detail the methodology employed in this study, while section 4 will present the results and discussion. Finally, section 5 will summarize the conclusions drawn from the research findings.

2. PROPOSED METHOD

According to de Souza [18], the Siamese Manhattan architecture enables the customization of networks for specific tasks, leveraging deep learning models like convolutional neural networks (CNN) and long short-term memory (LSTM). Recent advancements in STS tasks using DL have been achieved through CNN and LSTM models, which analyze words and sentences to capture both meaning and structural aspects for STS calculation [19]. Several studies have explored the implementation of Siamese Manhattan architecture in DL models, such as the research by Ranasinghe *et al.* [15], which proposes the implementation of Siamese Manhattan on several recurrent neuron network (RNN)-based models, including LSTM to calculate STS between text pairs in SemEval data. Their results showed that their model performed better than the Siamese neural network model first proposed by Mueller and Thyagarajan [14].

Zheng *et al.* [20] introduced a CNN-based model to detect semantic similarities in medical imaging reports (*e.g.*, ultrasound, MRI) and pathology. They enhanced the model's semantic understanding through embedding techniques, outperforming traditional approaches like keyword mapping, LSA, latent Dirichlet allocation (LDA), and Siamese LSTM. Similarly, Shi *et al.* [17] employed Siamese CNN to evaluate sentence similarity in Chinese, comparing Cosine and Manhattan similarity metrics. Their findings indicate that the Manhattan similarity metric outperforms other metrics, highlighting its effectiveness in this context.

Tran *et al.* [21] utilized a Siamese neural network architecture augmented with semantic features extracted from a knowledge graph, termed Siamese KG-LSTM, in conjunction with BioWordVec embeddings, to predict synonymous and non-synonymous pairs of biomedical terms within The unified medical language system (UMLS). The UMLS, developed by the U.S. National Library of Medicine, enhances computer systems' biomedical and health language comprehension. Their study demonstrated impressive performance metrics: 98.23% accuracy, 98.37% recall, 97.40% precision, and an F1-score of 96.41% for synonym and non-synonym prediction. Li and He [22] constructed an RNN model using a Siamese neural network architecture with Word2Vec for processing word vectors. They evaluated semantic similarity on a dataset of 22,655 pairs of ethnic medical questions using Euclidean, Cosine, and Manhattan distances. The study found that the

Manhattan distance achieved the highest similarity (F1-score of 97.34%), followed by Cosine (96.93%) and Euclidean distances (94.13%).

A recent study by Yang *et al.* [23] employed a Siamese neural network–CNN to predict drug-drug interaction (DDI) events by treating each drug separately with two CNN sub-networks sharing parameters to learn multimodal drug information. Merged feature representations were fed into a multilayer perceptron, enabling accurate categorization of DDI events from multimodal data.

Previous studies on the Siamese Manhattan model have focused on its implementation within single-model architectures such as CNN or LSTM. In contrast, our research explores its application in hybrid CNN-LSTM and LSTM-CNN models. This integration represents a novel contribution to our study. Furthermore, we introduce the adaptation of the Siamese Manhattan architecture with BioWordVec and a dual labelling process using CS and WMD, specifically for the biomedical domain, which addresses the challenges of contextual meaning and word order in biomedical texts.

3. METHOD

This study introduces CNN and LSTM models using the Siamese Manhattan architecture. Furthermore, we suggest combining these models CNN-LSTM and LSTM-CNN by integrating the Siamese Manhattan approach to model STS in medical reports. Several studies [24]–[26] show that this hybrid approach harnesses CNN and LSTM strengths to enhance STS accuracy and precision.

The dataset used in this research is the GENIA Biomedical event train data, which comprises unlabeled medical reports detailing the effects of drug use on individuals aimed at identifying specific medical conditions. This dataset includes 4,957 records with four variables: sentence, TriggerWord, TriggerWordLoc, and EventType. However, this study focuses on three variables: sentence (biomedical text), TriggerWord (trigger word in a sentence), and EventType (type of biomedical event related to the TriggerWord). Each pair: sentence-TriggerWord, sentence-EventType, and TriggerWord-EventType, will be analyzed for semantic similarity, resulting in 4,957 pairs for each combination. Figure 1 illustrates the research methodology.

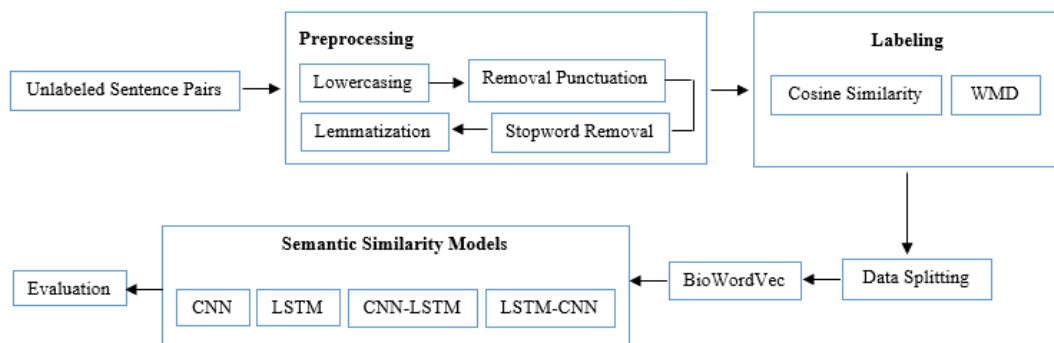


Figure 1. Research methodology

3.1. Preprocessing

A medical report is an official document authored by medical professionals that contains comprehensive details about a patient's diagnosis, treatment, and therapy. However, many medical reports suffer from poor sentence grammar [27], unstructured formatting [28], and large file sizes. Consequently, preprocessing is essential to facilitate more efficient analysis. The text preprocessing steps applied to these medical reports include:

- Lowercasing:** Converting all uppercase letters to lowercase is crucial because computer programs interpret variations in capitalization (e.g., Leukemia vs leukemia) as distinct word vectors, leading to different results [29].
- Remove punctuation:** Punctuation marks such as periods, commas, exclamation marks, and question marks are removed because computer programs do not comprehend punctuation, and their presence in the text is treated as noise.
- Stop word removal:** This step eliminates meaningless words, biomedical domain-specific words, or commonplace words in the text. Additionally, rare words are removed as they are too numerous and do not contribute significantly to STS tasks.
- Lemmatization:** This crucial step transforms words into their base morphological forms (lemmas), enhancing the consistency and accuracy of text analysis [30].

After undergoing various steps and transformations to ensure its cleanliness, structure, and readiness for further analysis, the initially 4,957 sentence pairs of textual data have now been reduced to only 1,409 pairs for each combination.

3.2. Labelling

STS tasks necessitate labelled data, thereby requiring the annotation of this data. The annotation process involves utilizing CS and WMD methods. CS calculates the similarity between two objects, represented by vectors of document keywords. The similarity score in CS ranges from 0 to 1: 0 signifies complete dissimilarity between the objects, whereas 1 indicates their identity or exact similarity [31]. As shown in (1) is employed to compute the similarity score using CS.

$$\text{Cosine similarity } (p, q) = \frac{(p \cdot q)}{\|p\| \cdot \|q\|} \quad (1)$$

where $\|p\|$ is length of vector p and $\|q\|$ is length of vector q

In contrast, WMD is a metric for quantifying the dissimilarity between two documents by computing the minimum distance required to convert one document's vocabulary into another [32]. The specific formulation of WMD is detailed in (2).

$$\text{WMD}(i, j) = \|x_i - x_j\|_2 \quad (2)$$

where WMD is word mover's distance and x_i is document weight.

The CS and WMD methods successfully labelled 1,409 sentence-TriggerWord pairs, 1,409 sentence-EventType pairs, and 1,409 TriggerWord-EventType pairs. An example demonstrating the results of this labelling process, explicitly using the CS method, can be found in Table 1. This table showcases how the CS method categorizes sentence-TriggerWord pairs, highlighting its effectiveness in labelling.

Table 1. Labelling using CS

Sentence	TriggerWord	Label
downregulation interferon regulatory factor 4 gene leukemic cell due hypermethylation CpG motif promoter region	downregulation	1
first treatment IRF4 negative lymphoid myeloid monocytic cell line methylation inhibitor deoxycytidine result time concentration dependent increase IRF4 mRNA protein level	negative	1
second use restriction PCR assay bisulfite sequencing identify specifically methylated CpG sit IRF4 negative IRF4 positive cell	negative	1
third clearly determine promoter methylation mechanism IRF4 downregulation via reporter gene assay detect association methylational status mRNA DNA methyltransferases Methyl-CpG-binding protein together data suggest CpG site specific irf4 promoter methylation putative mechanism downregulated IRF4 leukemia	downregulation	1
	downregulated	1

3.3. Data splitting

At this stage, the data is divided into training and testing sets, a process known as data splitting. The training data is used to build learning models, while the testing data is used to evaluate model performance [33]. Consequently, the proportion of training data must be more significant than that of testing data to prevent overfitting. The data splitting stage also includes the data division into the 'left' and 'right' input sides to align with the input requirements of the Siamese network.

This study's data is divided, with 75% allocated for training and 25% for testing. This 75/25 split is a standard approach in machine learning. It is balancing the need for comprehensive training with enough data to test and validate the model's performance rigorously.

3.4. BioWordVec

BioWordVec, developed by Zhang *et al.* [34], is integrated into our model as input embedding layers to analyze the specific and local context of words within medical reports. This word embedding is trained using biomedical corpora from biomedical literature and medical subject headings (MeSH) domain knowledge. Zhang *et al.* [34] utilized a subword embedding model to improve understanding of text sequences and medical terminology in MeSH, enhancing biomedical word representations and semantic comprehension.

BioWordVec is categorized into intrinsic and extrinsic types. Intrinsic BioWordVec is commonly used to predict semantic similarity among words, terms, or sentences. Conversely, extrinsic BioWordVec serves as feature input in various NLP tasks, such as relation extraction and text classification. This dual

functionality makes BioWordVec a versatile tool for boosting the performance of NLP applications in the biomedical field.

3.5. Semantic similarity models

3.5.1. Convolutional neural networks (CNN) and long short-term memory (LSTM)

A CNN DL algorithm identifies sentence patterns and semantic relationships through a convolutional layer linked to local features. Typically, a CNN consists of multiple layers: input, convolutional, pooling, fully connected, and output layers [35]. CNN has succeeded in various NLP tasks, such as sentence representation, search query retrieval, and semantic parsing. They are particularly effective in identifying semantic similarities between text pairs, excelling in tasks related to STS [20]. However, CNNs have a limitation: their architecture does not account for word order relationships in sentences, which means they cannot analyze the sequence of words.

3.5.2. Long short-term memory (LSTM)

Unlike CNN, the LSTM architecture is specifically designed for processing sequence data. LSTM efficiently discards irrelevant information and retains only essential information, sequentially capturing the essence of sentences through cell states and gates [36]. This capability makes LSTM particularly well-suited for tasks involving semantic similarity, as they can understand the context and meaning of words in sentences sequentially.

3.5.3. Hybrid model

The hybrid model analyzes words from both general and local contexts. Terms are interpreted using word embedding for general context and specific semantic and syntactic features for local context. There are two hybrid models: CNN-LSTM and LSTM-CNN. Hybrid CNN-LSTM and LSTM-CNN models process data in different sequences. In the hybrid CNN-LSTM model, CNN first extracts the local context of each word in a sentence, followed by LSTM to check word order. Conversely, in the hybrid LSTM-CNN model, LSTM initially checks the word order, and then CNN extracts the local context [19].

3.5.4. Hidden layer

Hidden layers are crucial for improving accuracy and managing time complexity in learning models. Research by Uzair and Jamil [37] indicates that three hidden layers offer optimal performance. However, using more than three hidden layers directly affects the model's accuracy negatively.

3.6. Evaluation

Model performance is evaluated using the confusion matrix [38], where accurate STS calculations appear along the diagonal. Errors in STS calculations are represented outside this diagonal line. This method provides a clear and structured evaluation of how well the model predicts textual similarities.

4. RESULTS AND DISCUSSION

This section presents the results achieved through the proposed method. Various pairs of sentences were evaluated for STS, including sentence-TriggerWord, sentence-EventType, and TriggerWord-EventType. The comparison of each model's runtime is illustrated in Figures 2(a) to 2(c) respectively.

Applying the Siamese Manhattan architecture to four different DL methods yielded identical results with 100% accuracy and validation accuracy. However, each model varies in the time required for STS calculations. CNN emerges as the most time-efficient model for training compared to others. Models like CNN-LSTM and LSTM-CNN hybrids also exhibit good efficiency. In contrast, the single LSTM model proves less efficient due to longer processing times. Determining the optimal model is challenging when considering accuracy, validation accuracy, and training duration alone.

An alternative method to evaluate model performance involves analyzing the loss graph. Loss graphs indicate whether a model is underfitting, well-fitted, or overfitting. Underfitting signifies poor performance during training or testing. A well-fitted model displays consistent performance across training and testing phases, whereas overfitting manifests as a significant gap between training and testing results. The initial model examined through loss graphs was the CNN model applied to sentence-TriggerWord pairs. Detailed comparisons of these graphs are depicted in Figures 3(a) to 3(d).

Overall, among CNN models evaluated for STS in sentence-TriggerWord pairs, the model using three hidden layers with the WMD labelling method stands out as the most suitable. This conclusion is drawn from the loss graph analysis, where both loss and validation loss values converge effectively. In contrast, other CNN models, as depicted in Figure 3, show signs of underfitting as they fail to converge. Integrating the WMD method with three hidden layers proves beneficial, enhancing overall model performance significantly.

The next model under evaluation in the loss graph is the CNN-LSTM hybrid model, illustrated in Figures 4(a) to 4(d). The loss graphs in Figures 4(a) and 4(b) illustrate that the CNN-LSTM hybrid model, utilizing two hidden layers with different labelling methods, experiences a decreasing loss value accompanied by an increasing validation loss value, indicating overfitting. Consequently, this model may not generalize well to new data for determining STS. Conversely, the CNN-LSTM hybrid models depicted in Figures 4(c) and 4(d) exhibit more balanced loss graphs, where the loss and validation loss values are closely aligned. However, even with these improvements, these models still fail to determine STS accurately. The last model analyzed through the loss graph is the LSTM-CNN hybrid model, displayed in Figures 5(a) to 5(d). Based on Figures 5(a) to 5(d), the loss and validation loss values of the LSTM-CNN hybrid model converge consistently across various labelling methods and hidden layer configurations. This convergence signifies that the model demonstrates robust performance and generalizability during training and when applied to new data.

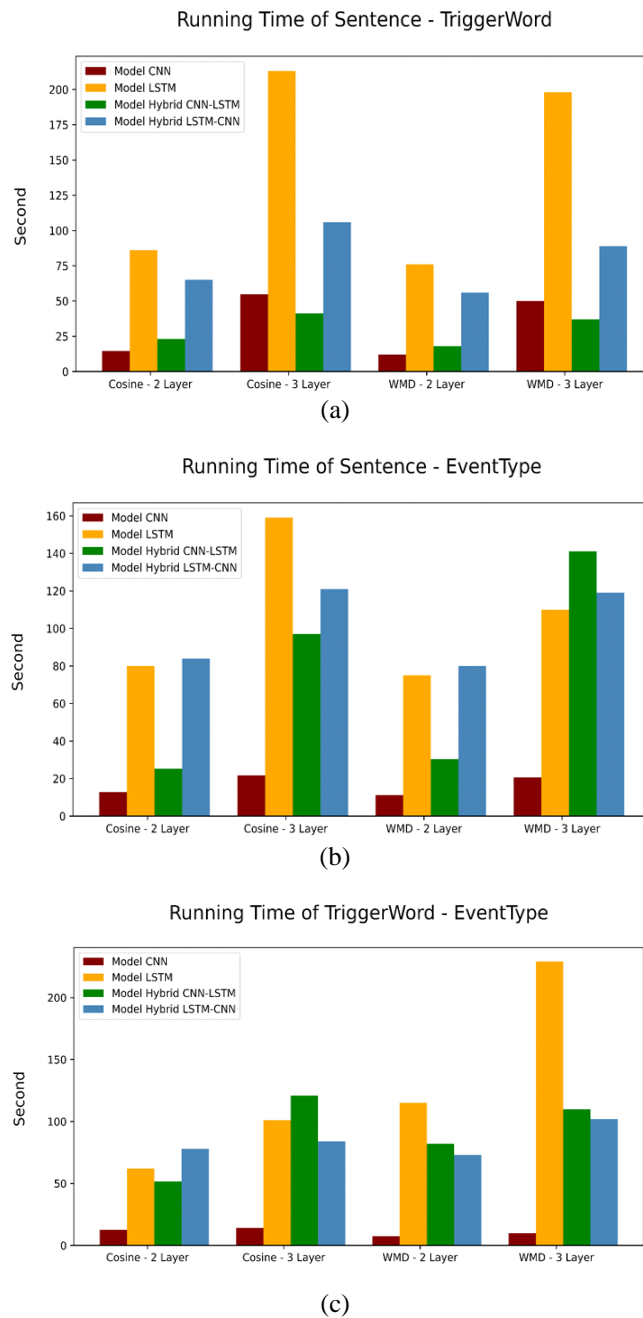


Figure 2. Comparison of running time sentence pairs on (a) sentence-TriggerWord, (b) sentence-EventType, and (c) TriggerWord-EventType

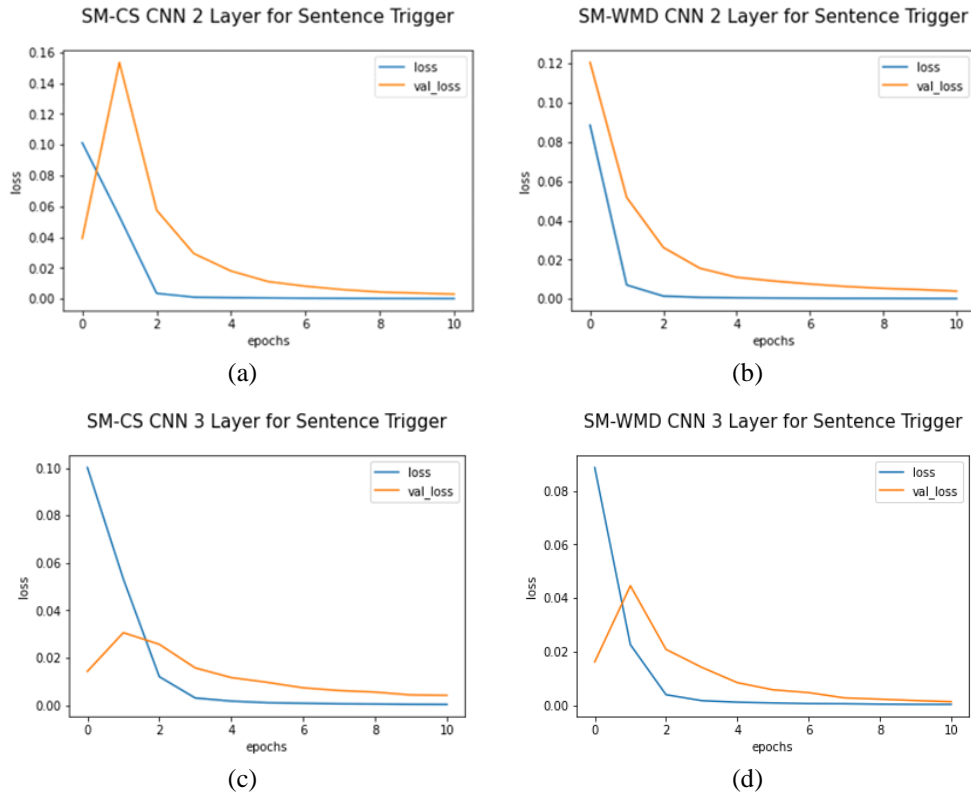


Figure 3. CNN model loss graph on sentence-TriggerWord using (a) 2 Layer+CS, (b) 2 Layer+WMD, (c) 3 Layer+CS, and (d) 3 Layer+WMD

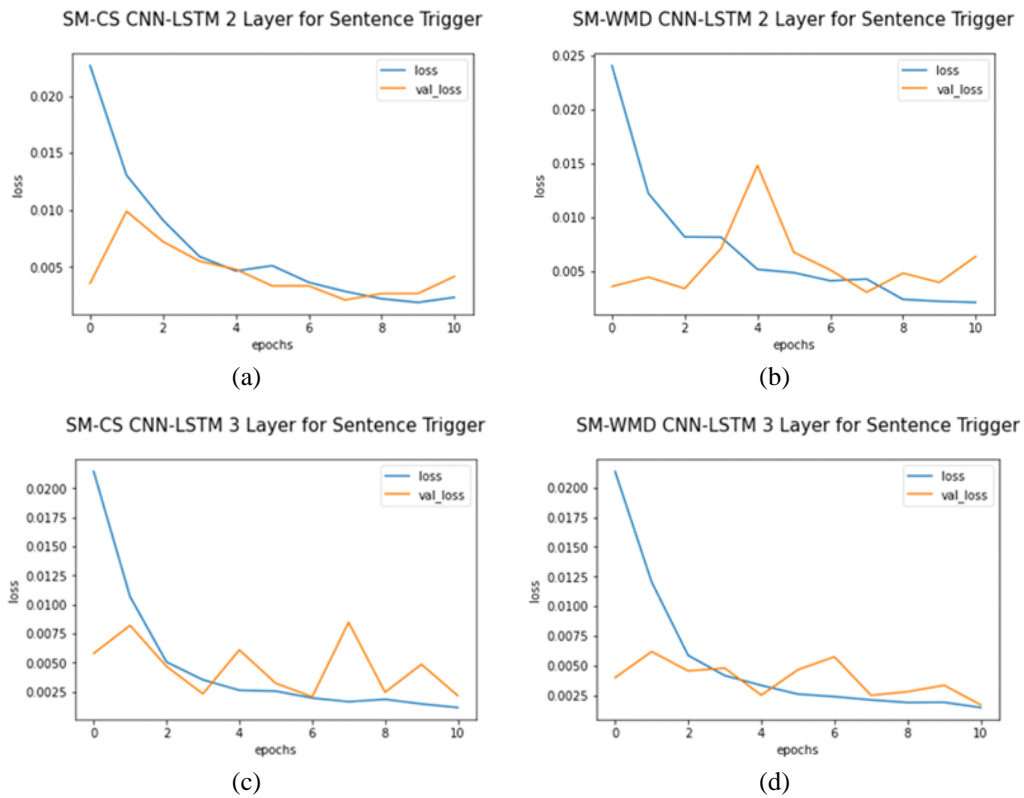


Figure 4. CNN-LSTM hybrid model loss graph on sentence-TriggerWord using (a) 2 Layer+CS, (b) 2 Layer+WMD, (c) 3 Layer+CS, and (d) 3 Layer+WMD

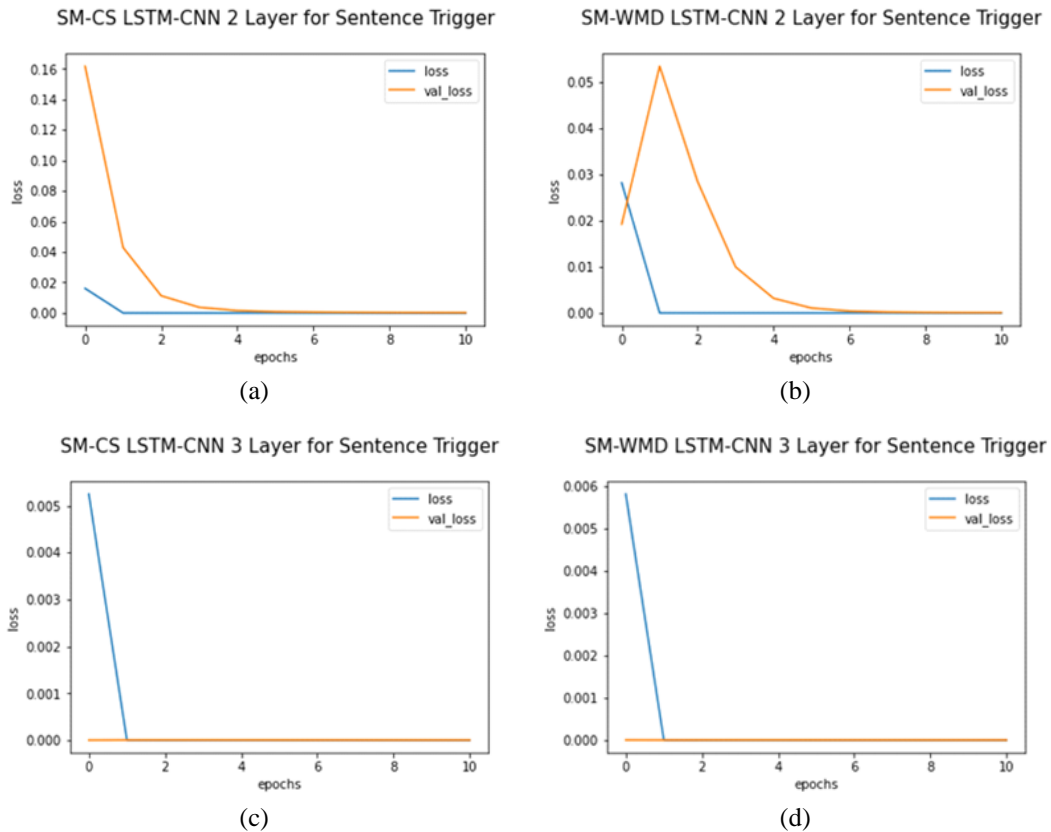


Figure 5. LSTM-CNN hybrid model loss graph on sentence-TriggerWord using (a) 2 Layer+CS, (b) 2 Layer+WMD, (c) 3 Layer+CS, and (d) 3 Layer+WMD

Based on the analysis of the loss graph, it is evident that more complex models with more than two hidden layers outperform those with only two layers. Furthermore, the choice of the labelling method significantly impacts the training time, with the WMD method proving superior to CS due to its faster convergence. In the LSTM-CNN hybrid model, the loss graph remains stable across different labelling methods and hidden layer configurations. Thus, the LSTM-CNN hybrid model is optimal for determining STS in sentence-TriggerWord, sentence-EventType, and EventType-TriggerWord pairs. These findings underscore the model's consistency and effectiveness across varied sentence pairings.

The STS calculation results from the LSTM-CNN hybrid model, identified as the most effective model for determining semantic similarities across the three sentence pairs, are presented in the confusion matrices shown in Figures 6(a) to 6(c). These matrices illustrate the performance of the LSTM-CNN hybrid model, which features a sophisticated architecture and utilizes the optimal labelling method, WMD. Figure 6(a) illustrates that the sentence-TriggerWord pair achieves a similarity score of 1, indicating identical semantic meaning between the sentence and TriggerWord variables. The sentence variable comprises sentences related to leukemia, encompassing genetic regulation, inter-protein interactions, genes or proteins, and associated biological processes. In contrast, the TriggerWord variable contains descriptive terms as indicators or keywords to identify specific biomedical events within the sentence and estimate their timing. Therefore, based on the STS calculations for the sentence-TriggerWord pair, the model effectively extracts meaningful information and crucial elements from biomedical texts pertinent to leukemia research.

Similarly, Figure 6(b) reveals that the sentence-EventType pair achieves a similarity score of 1, indicating identical semantic meaning between these variables. Alongside matching the TriggerWord's semantic meaning, the sentence variable aligns closely with the EventType variable, which denotes specific biological activities within the context of leukemia. STS calculations in these pairs facilitate the identification of correlations between various biomedical events and leukemia, aiding researchers and medical professionals. Furthermore, analyzing semantic relationships between these biomedical events and texts enhances understanding of how leukemia treatments influence specific biological pathways and related biomedical phenomena. This approach contributes deeper insights into leukemia therapy and its implications for biomedical research and treatment strategies. Lastly, based on Figure 6(c), the TriggerWord-EventType pair

achieves a similarity score 1, indicating a solid semantic relationship between all trigger words in the TriggerWord variable and the biomedical event types in the EventType variable. This analysis reveals significant insights into the biological mechanisms associated with leukemia disease.

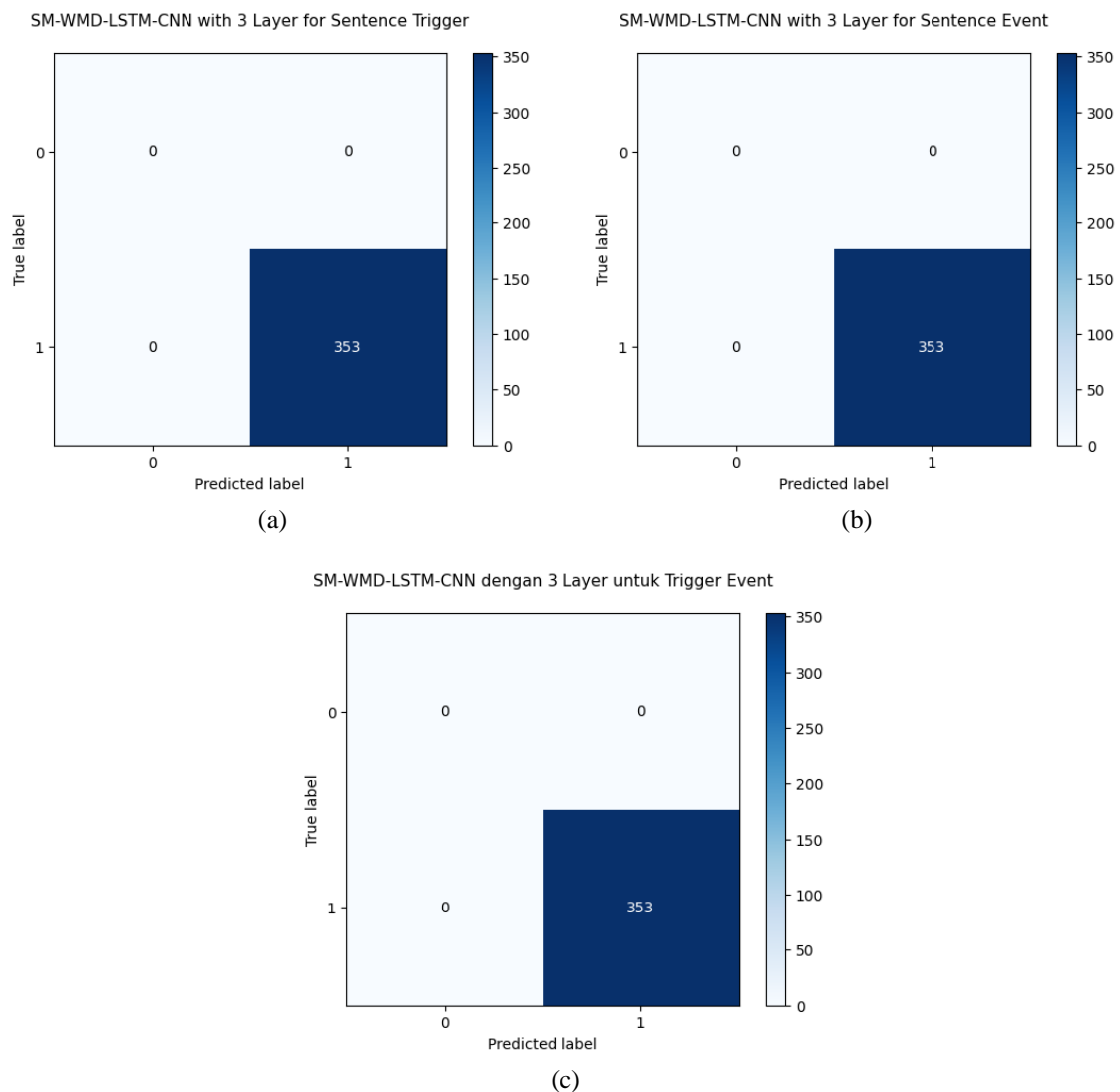


Figure 6. Confusion matrix on (a) sentence – TriggerWord, (b) sentence – EventType, and (c) TriggerWord – EventType

The top-performing model in this research study is benchmarked against CNN, LSTM, and hybrid CNN-LSTM models and models proposed in previous studies. However, it is worth noting that prior studies often did not exclusively focus on the biomedical domain due to its relatively limited research volume compared to other fields. This benchmarking is presented in Table 2.

The proposed method exhibits superior performance compared to previous models, as evidenced by the comparative results presented in Table 2. Among these, Tran *et al.* [21] achieved notable success using BioWordVec to predict synonyms and non-synonyms in the biomedical domain with an accuracy exceeding 98%. This study, alongside Tran *et al.* [21], underscores the effectiveness of employing pre-trained word embeddings tailored to specific research domains, such as BioWordVec. Specifically trained on datasets from PubMed and MeSH, BioWordVec enhances the STS model's performance by providing domain-specific embeddings relevant to the biomedical data used in this research. Moreover, the high accuracy achieved in this study is also attributable to the selection of appropriate similarity metrics, particularly highlighted in Table 2, where the Manhattan metric is shown to potentially improve the STS model's accuracy significantly.

Table 2. Benchmarking analysis of STS calculations

Title	Authors	Methods	Performance
Unlabeled short text similarity with LSTM encoder	Yao <i>et al.</i> [39]	a. Data: MSR paraphrase and Quora dataset (Unlabeled) b. Labelling: cosine similarity c. Domain: general d. Word embedding: Word2Vec Model: LSTM encoder	Accuracy MSR: 88,00% Recall MSR: 87,00% Accuracy Quora: 85,00% Recall Quora: 92,00%
STS with Siamese neural networks	Ranasinghe <i>et al.</i> [15]	a. Data: SemEval 2017 dataset (labeled) b. Domain: biomedical c. Word embedding: custom-trained embedding Model: Siamese Manhattan – LSTM	Accuracy: 86,51%
Detection of medical text semantic similarity based on convolutional neural network	Zheng <i>et al.</i> [20]	a. Data: imaging and pathology report-pairs (labeled) b. Domain: biomedical c. Word embedding: CMESH Model: Siamese neural network – CNN using LIME algorithm for similarity output	Recall: 93,70% Precision: 94,50% F1-Score: 94,10%
A Siamese CNN architecture for learning Chinese sentence similarity	Shi <i>et al.</i> [17]	a. Data: Chinese sentence pairs (labeled) b. Domain: general c. Word embedding: custom-trained embedding Model: Siamese neural network – CNN with two different metric distances (Cosine and Manhattan)	Accuracy Cosine: 77,05% Accuracy Manhattan: 77,31%
Siamese KG – LSTM: a deep learning model for enriching UMLS Meta thesaurus synonymy	Tran <i>et al.</i> [21]	a. Data: the UMLS dataset (labeled) b. Domain: biomedis c. Word embedding: BioWordVec Model: Siamese Manhattan - LSTM with knowledge graph	Accuracy: 98,23% Recall: 93,87% Precision: 97,86% F1-Score: 96,41%
Similarity matching of medical questions based on Siamese network	Li and He [22]	a. Data: ethnic medical Question dataset (labeled) b. Domain: Biomedis c. Word embedding: Word2Vec d. Model: Siamese neural network – RNN with Manhattan distance	F1-Score Euclidean: 94,13% F1-Score Cosine: 96,93% F1-Score Manhattan: 97,34%
Proposed method	Kurniasari <i>et al.</i>	a. Data: GENIA event dataset (unlabeled) b. Labeling: CS and WMD c. Word embedding: BioWordVec Model: Siamese Manhattan-hybrid LSTM CNN	Accuracy: 100% Recall: 100% Precision: 100% F1-Score: 100%

5. CONCLUSION

This research identifies the optimal model for detecting semantic similarity among leukemia-related biomedical terms, a hybrid LSTM-CNN model combining Siamese Manhattan and LSTM-CNN architectures. Results indicate superior performance compared to models proposed in previous research despite differing domains between studies. The research aims to aid researchers and medical professionals in uncovering correlations between various biomedical events and leukemia, understanding semantic connections between trigger words and biomedical events, and automatically extracting critical information from biomedical texts to enhance leukemia-related information retrieval. Moreover, the study seeks to deepen understanding of how leukemia treatments impact specific biological pathways and subsequent biomedical processes. Future research directions could explore further advancements in the STS task by extracting relationships between biomedical entities, thereby enhancing insights into leukemia and its biomedical implications.

ACKNOWLEDGEMENTS

This research received funding from a Faculty of Mathematics and Natural Sciences grant for the 2023 academic year.

REFERENCES




- [1] A. Chaves, C. Kesiku, and B. Garcia-Zapirain, "Automatic text summarization of biomedical text data: A systematic review," *Information*, vol. 13, no. 8, Aug. 2022, doi: 10.3390/info13080393.
- [2] Y. Wang, S. Fu, F. Shen, S. Henry, O. Uzuner, and H. Liu, "The 2019 n2c2/OHNL track on clinical semantic textual similarity: Overview," *JMIR Medical Informatics*, vol. 8, no. 11, Nov. 2020, doi: 10.2196/23375.
- [3] F. Alam, M. Afzal, and K. M. Malik, "Comparative analysis of semantic similarity techniques for medical text," in *2020 International Conference on Information Networking (ICOIN)*, Jan. 2020, vol. 2020-Janua, pp. 106–109, doi: 10.1109/ICOIN48656.2020.9016574.
- [4] Q. Chen, J. Du, S. Kim, W. J. Wilbur, and Z. Lu, "Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records," *BMC Medical Informatics and Decision Making*, vol. 20, no. S1, Apr. 2020, doi: 10.1186/s12911-020-1044-0.

- [5] M. Li, X. Zhou, K. H. Ryu, and N. Theera-Umpon, "An ensemble semantic textual similarity measure based on multiple evidences for biomedical documents," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–14, Aug. 2022, doi: 10.1155/2022/8238432.
- [6] S. P. and A. P. Shaji, "A survey on semantic similarity," in *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, Dec. 2019, pp. 1–8, doi: 10.1109/ICAC347590.2019.9036843.
- [7] W. Zhao, X. Liu, J. Jing, and R. Xi, "Re-LSTM: A long short-term memory network text similarity algorithm based on weighted word embedding," *Connection Science*, vol. 34, no. 1, pp. 2652–2670, 2022, doi: 10.1080/09540091.2022.2140122.
- [8] X. Yang, X. He, H. Zhang, Y. Ma, J. Bian, and Y. Wu, "Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models," *JMIR Medical Informatics*, vol. 8, no. 11, Nov. 2020, doi: 10.2196/19735.
- [9] Y. Wang *et al.*, "MedSTS: a resource for clinical semantic textual similarity," *Language Resources and Evaluation*, vol. 54, no. 1, pp. 57–72, Mar. 2020, doi: 10.1007/s10579-018-9431-1.
- [10] A. Romanov and C. Shivade, "Lessons from natural language inference in the clinical domain," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1586–1596, doi: 10.18653/v1/D18-1187.
- [11] Y. Wang, K. Verspoor, and T. Baldwin, "Learning from unlabelled data for clinical semantic textual similarity," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 2020, pp. 227–233, doi: 10.18653/v1/2020.clinicalnlp-1.25.
- [12] M. Brunila and J. LaViolette, "WMDecompose: A framework for leveraging the interpretable properties of word mover's distance in sociocultural analysis," in *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2021, pp. 154–167, doi: 10.18653/v1/2021.latechclfl-1.18.
- [13] H. Yamagiwa, S. Yokoi, and H. Shimodaira, "Improving word mover's distance by leveraging self-attention matrix," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Nov. 2023, pp. 11160–11183, doi: 10.18653/v1/2023.findings-emnlp.746.
- [14] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, pp. 2786–2792, Mar. 2016, doi: 10.1609/aaai.v30i1.10350.
- [15] T. Ranasinghe, C. Orăsan, and R. Mitkov, "Semantic textual similarity with Siamese neural networks," in *Proceedings - Natural Language Processing in a Deep Learning World*, Oct. 2019, vol. 2019-Septe, pp. 1004–1011, doi: 10.26615/978-954-452-056-4_116.
- [16] D. Chicco, "Siamese neural networks: An overview," in *Methods in Molecular Biology*, vol. 2190, 2021, pp. 73–94, doi: 10.1007/978-1-0716-0826-5_3.
- [17] H. Shi, C. Wang, and T. Sakai, "A Siamese CNN architecture for learning Chinese sentence similarity," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, 2020, pp. 24–29.
- [18] J. V. A. de Souza, L. E. S. E. Oliveira, Y. B. Gumiel, D. R. Carvalho, and C. M. C. Moro, "Exploiting Siamese neural networks on short text similarity tasks for multiple domains and languages," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12037 LNAI, 2020, pp. 357–367, doi: 10.1007/978-3-030-41505-1_34.
- [19] E. L. Pontes, S. Huet, A. C. Linhares, and J.-M. Torres-Moreno, "Predicting the semantic textual similarity with Siamese CNN and LSTM," *Prepr. arXiv.810.10641*, 2018.
- [20] T. Zheng *et al.*, "Detection of medical text semantic similarity based on convolutional neural network," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12911-019-0880-2.
- [21] T. T. T. Tran *et al.*, "Siamese KG-LSTM: A deep learning model for enriching UMLS metathesaurus synonymy," in *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, Nov. 2020, pp. 281–286, doi: 10.1109/KSE50997.2020.9287797.
- [22] Q. Li and S. He, "Similarity matching of medical question based on Siamese network," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, Apr. 2023, doi: 10.1186/s12911-023-02161-z.
- [23] Z. Yang, K. Tong, S. Jin, S. Wang, C. Yang, and F. Jiang, "CNN-Siam: Multimodal Siamese CNN-based deep learning approach for drug-drug interaction prediction," *BMC Bioinformatics*, vol. 24, no. 1, Mar. 2023, doi: 10.1186/s12859-023-05242-y.
- [24] T. Liu, J. Bao, J. Wang, and Y. Zhang, "A Hybrid CNN-LSTM algorithm for online defect recognition of CO2 welding," *Sensors*, vol. 18, no. 12, Dec. 2018, doi: 10.3390/s18124369.
- [25] M. Mansoor, Z. ur Rehman, M. Shaheen, M. A. Khan, and M. Habib, "Deep learning based semantic similarity detection using text data," *Information Technology and Control*, vol. 49, no. 4, pp. 495–510, Dec. 2020, doi: 10.5755/j01.itc.49.4.27118.
- [26] R. Beniwal, D. Bhardwaj, B. P. Raghav, and D. Negi, "Text similarity identification based on CNN and CNN-LSTM model," in *Second International Conference on Sustainable Technologies for Computational Intelligence*, 2022, pp. 47–58, doi: 10.1007/978-981-16-4641-6_5.
- [27] H. S. Yahia and A. M. Abdulazeez, "Medical text classification based on convolutional neural network: a review," *International Journal of Science and Business*, vol. 5, no. 3, pp. 27–41, 2021.
- [28] H.-J. Kong, "Managing unstructured big data in healthcare system," *Healthcare Informatics Research*, vol. 25, no. 1, pp. 1–2, 2019, doi: 10.4258/hir.2019.25.1.1.
- [29] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 40–46, doi: 10.18653/v1/W18-5406.
- [30] I. Akhmetov, A. Pak, I. Ualiyeva, and A. Gelbukh, "Highly language-independent word lemmatization using a machine-learning classifier," *Computación y Sistemas*, vol. 24, no. 3, pp. 1353–1364, Sep. 2020, doi: 10.13053/cys-24-3-3775.
- [31] A. Apriani, H. Zakiyudin, and K. Marzuki, "Application of the cosine similarity algorithm and weighting of the TF-IDF system for new student admissions on private campuses," (in Bahasa), *Jurnal Bumigora Information Technology (BITE)*, vol. 3, no. 1, pp. 19–27, Jul. 2021, doi: 10.30812/bite.v3i1.1110.
- [32] N. Pribadi, R. Sarno, A. Ahmadiyah, and K. Sungkono, "Semantic recommender system based on semantic similarity using fast text and word mover's distance," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 2, pp. 377–385, Apr. 2021, doi: 10.22266/ijies2021.0430.34.
- [33] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of Soil," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–15, Feb. 2021, doi: 10.1155/2021/4832864.
- [34] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, "BioWordVec, improving biomedical word embeddings with subword information and MeSH," *Scientific Data*, vol. 6, no. 1, May 2019, doi: 10.1038/s41597-019-0055-0.
- [35] W. K. H. W. K. Amir, A. B. M. Soom, A. M. Jasin, J. Ismail, A. Asmat, and R. A. R. An, "Sales forecasting using convolution neural network," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 30, no. 3, pp. 290–301, May 2023, doi: 10.37934/araset.30.3.290301.




- [36] S. M. Al-Selwi, M. F. Hassan, S. J. Abdulkadir, and A. Muneer, "LSTM inefficiency in long-term dependencies regression problems," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 30, no. 3, pp. 16–31, May 2023, doi: 10.37934/araset.30.3.1631.
- [37] M. Uzair and N. Jamil, "Effects of hidden layers on the efficiency of neural networks," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, Nov. 2020, pp. 1–6, doi: 10.1109/INMIC50486.2020.9318195.
- [38] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Diagnosing of disease using machine learning," in *Machine Learning and the Internet of Medical Things in Healthcare*, Elsevier, 2021, pp. 89–111, doi: 10.1016/B978-0-12-821229-5.00003-3.
- [39] L. Yao, Z. Pan, and H. Ning, "Unlabeled short text similarity with LSTM encoder," *IEEE Access*, vol. 7, pp. 3430–3437, 2019, doi: 10.1109/ACCESS.2018.2885698.

BIOGRAPHIES OF AUTHORS






Dian Kurniasari    received her bachelor's degree in statistics from Gadjah Mada University, Indonesia, in 1994, followed by her M.Sc. in applied mathematics from Curtin University of Technology in 2002. Currently, Dian Kurniasari is an associate professor in the Department of Mathematics at the Faculty of Mathematics and Natural Sciences, University of Lampung, where her academic pursuits primarily focus on statistics and data science. For further inquiries, she can be reached at email: dian.kurniasari@fmipa.unila.ac.id.






Mustofa Usman    received his bachelor's degree from the University of Lampung, Indonesia, in 1983. He earned his M.A. from the State University of New York, Albany, in 1988 and completed his Ph.D. at Kansas State University in 1995. Currently, he holds the position of professor in the Department of Mathematics within the Faculty of Mathematics and Natural Sciences at the University of Lampung. His academic pursuits focus on linear models, time series modeling, and applied statistics. For further inquiries, he can be reached at email: usman_alpha@yahoo.com.



Warsono    received his bachelor's and master's degrees from the Bogor Agricultural Institute, Indonesia, in 1985 and 1991, respectively. In 2000, he completed his PhD at the University of Alabama, Birmingham, United States. Currently, Warsono is an associate professor in the Department of Mathematics in the Faculty of Mathematics and Natural Sciences at the University of Lampung. His research focuses on statistical modeling, theoretical distribution, and applied statistics. For further inquiries, he can be reached at email: warsono.1963@fmipa.unila.ac.id.



Favorisen Rosyking Lumbanraja    received his bachelor's and master's degrees from Bogor Agricultural Institute, Indonesia, in 2007 and 2011, respectively. He completed his Ph.D. at Kanazawa University in 2017. Currently, he serves as an assistant professor in the Department of Computer Science at the Faculty of Mathematics and Natural Sciences, University of Lampung. His research focuses on bioinformatics, computing in mathematics, natural sciences, engineering, and medicine. For inquiries, he can be reached at email: favorisen.lumbanraja@fmipa.unila.ac.id.