

A Semantic Text Similarity Model for Predicting Medical Transcription Data

Daniel A. Folorunso ^{1,✉}, Adio T. Akinwale ¹, Alaba O. Adejimi ¹, Oluwaseun I. Adebisi ²

1. Computer Science, Federal University of Agriculture, Abeokuta, NGA

2. Electrical and Electronics Engineering, Federal University of Agriculture, Abeokuta, NGA

Received: April 05, 2026 | Review began: April 07, 2026 | Review ended: April 14, 2026 | Published: April 28, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Background

The health sector faces challenges in analysing unstructured medical transcription data, particularly in identifying semantic similarities between clinical question pairs for information retrieval. A major challenge is that it is not feasible to obtain sufficiently large and representative data for specialised machine learning models due to privacy policies. Data augmentation could help alleviate these challenges and, therefore, requires investigation.

Methods

This study investigated two models, Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) and Artificial Neural Network (ANN), for semantic text similarity classification. An annotated dataset of 3,048 medical question pairs from Hugging Face was employed. The LSTM-RNN model was developed using an Embedding layer and an LSTM layer, whereas the ANN model was developed without these layers. Experiments were conducted in TensorFlow for 100 epochs with an 80:20 train-validation split.

Results

Six performance metrics were recorded: Accuracy, Binary Cross-Entropy Loss, Area Under the Curve (AUC), Precision, Recall, and F1-Score. The augmented LSTM-RNN significantly outperformed other configurations, achieving a validation accuracy of 95.31%, a loss of 0.2931, an AUC of 97.39%, a precision of 94.69%, a recall of 96.21 %, and an F1-score of 95.44%. Without augmentation, the LSTM-RNN validation accuracy dropped to 58.03%. The augmented ANN achieved a validation accuracy of 82.87%, while the non-augmented ANN struggled with 50.49% accuracy.

Conclusions

The inclusion of LSTM and Embedding layers allowed the LSTM-RNN to capture contextual dependencies that the ANN could not. The results demonstrate that data augmentation is important for achieving high-performance metrics in clinical text analysis, where data is limited.

Categories: Health Informatics, Deep Learning, Natural Language Processing (NLP)

Keywords: natural language processing, clinical text analysis, data augmentation, artificial neural network, long short-term memory, recurrent neural network, health informatics, medical transcription, logistic regression analysis, semantic text similarity

How to cite this article:

Folorunso D A, Akinwale A T, Adejimi A O, et al. (April 28, 2026) A Semantic Text Similarity Model for Predicting Medical Transcription Data. Cureus J Comput Sci 3 : es44389-026-00062-6. DOI <https://doi.org/10.7759/s44389-026-00062-6>

Introduction

The application of Artificial Intelligence and Machine Learning in the health sector has seen a significant rise, with large language models recently enhancing efficiency in handling medical records and reports [1]. However, traditional medical transcription often suffers from errors stemming from physician dictation proficiency or the experience levels of transcriptionists and clinical coders [2]. Such inaccuracies in medication documentation and clinical coding can lead to significant potential for medical error [3,4].

As the healthcare industry generates vast volumes of data, there is an increasing need for scalable and accurate analytic tools to process electronic medical records [5]. Semantic Text Similarity (STS) is the contextual relationship between a pair of texts, often depicted with a similarity score, and it is further established that vector embedding-based approaches have significantly improved STS analysis performance, making these approaches adoptable for Clinical STS tasks [6]. While techniques like named entity recognition show potential in clinical settings [7], Recurrent Neural Networks (RNNs) remain a standard for sequence-based text analysis [8]. Specifically, various architecture types of RNNs are now being utilised to handle the complexities of medical data [9]. This study explores the use of Long Short-Term Memory (LSTM)-RNN to analyse medical questions and predict similarity scores, addressing the challenge of data scarcity caused by privacy regulations [10].

Materials And Methods

The implementation of the proposed models was carried out using Python 3.9 within the PyCharm development environment. Key libraries utilised include NumPy and Pandas for data manipulation, Natural Language Toolkit (NLTK) for text preprocessing, and TensorFlow for model development and training.

The dataset used in this study consists of 3,048 annotated medical question pairs obtained from the Hugging Face repository [11]. Each pair is labelled to indicate whether the questions are semantically similar or dissimilar, as shown in Table 7. Before model development, the dataset underwent a structured preprocessing pipeline to enhance data quality and model performance.

dr_id	Question_1	Question_2	Label
1	After how many hour from drinking an antibiotic can I drink alcohol?	I have a party tonight and I took my last dose of Azithromycin this morning. Can I have a few drinks?	1
1	After how many hour from drinking an antibiotic can I drink alcohol?	I vomited this morning and I am not sure if it is the side effect of my antibiotic or the alcohol I took last night...	0
1	Am I over weight (192.9) for my age (39)?	I am a 39 y/o male currently weighing about 193 lbs. Do you think I am overweight?	1

TABLE 1: Dataset Sample

Initially, irrelevant attributes such as the *dr_id* column were dropped to retain only cogent textual features. To address the limited dataset size and potential class imbalance, oversampling techniques were applied using the Pandas *DataFrame.loc* function to expand the existing data by duplication, followed by data augmentation using a random word synonym substitution approach within the *fast_aug* package with a substitution probability of 5%. This step was necessary to improve model generalisation while preserving semantic integrity.

The English words were subsequently reduced to their base forms using NLTK's *Lemmatization* and *corpus* modules, with a focus on noun tokens to emphasise entity-level information within the dataset. This was intended to reduce lexical variability while preserving core semantic content. Similarly, stopwords were sourced from the NLTK corpus and removed alongside words or characters that were not strictly alphabetic from the texts to reduce noise in the resulting data.

The processed text data were then transformed into numerical representations using TensorFlow's *TextVectorization* layer with a vocabulary size capped at 300 tokens, corresponding to the most frequent words in the dataset. This constraint was introduced to ensure a fixed feature space across all experimental variants and to mitigate sparsity in this low-resource, domain-specific dataset.

While the augmentation process increased lexical variability, the dataset remains relatively small and domain-specific, which informed the decision to maintain a fixed vocabulary size to stabilise model training under constrained data conditions.

To ensure uniform input dimensions and avoid ragged tensor structures due to varying sequence lengths, zero-padding was applied in NumPy, extending each sequence to the maximum sequence length observed in the dataset by appending a zero matrix.

After the initial oversampling and preprocessing phases, a *drop_duplicates* function from Pandas *DataFrame* was executed to ensure the model did not memorise identical strings. To maintain statistical integrity and prevent class bias, the dataset was grouped by label, and each group was sampled based on the minimum class count. This resulted in a balanced, unique dataset of 15,240 samples. This rigorous sampling ensured that the results reflect the model's ability to generalise across varied medical question pairs rather than simply identifying duplicated entries.

For model input preparation, the vectorised representations of each question pair were concatenated to form a single feature vector via the *Flatten* layer in TensorFlow. Two distinct architectures were developed for comparative analysis. The first model is an LSTM-RNN, which incorporates an *Embedding* layer to generate dense vector representations and an *LSTM* layer to capture sequential dependencies within the text, as shown in the flow in Figure 1 [12]. The second model is a baseline Artificial Neural Network (ANN), which operates on the same input data but excludes both embedding and sequential processing layers, relying instead on fully connected dense layers. Both models included a *BatchNormalization* layer to aid vectors with outliers in terms of stability and training speed. For the ANN, the normalisation came after the process of concatenation, while it was added after the embedding layer in the LSTM-RNN model, as the embedding layer expects *int* data that serves as a "primary key" for the generated vector representations, whereas the *BatchNormalization* returns *float* data.

The embedding layer was omitted from the baseline ANN to avoid introducing extraneous noise into a non-sequential architecture. While an embedding layer generates dense vector representations, its primary utility in this study is to facilitate the LSTM's ability to capture combinatory and sequential dependencies. In a standard ANN, which relies on fully connected dense layers without a recurrent mechanism, adding an embedding layer often results in the model merely checking for the existence of values rather than understanding the "time series" or order of the patient's transcripts. Therefore, this architectural gap provides a controlled baseline highlighting the importance of sequential processing in clinical semantic similarity.

Both models were trained and evaluated under identical conditions to ensure a fair comparison of performance.

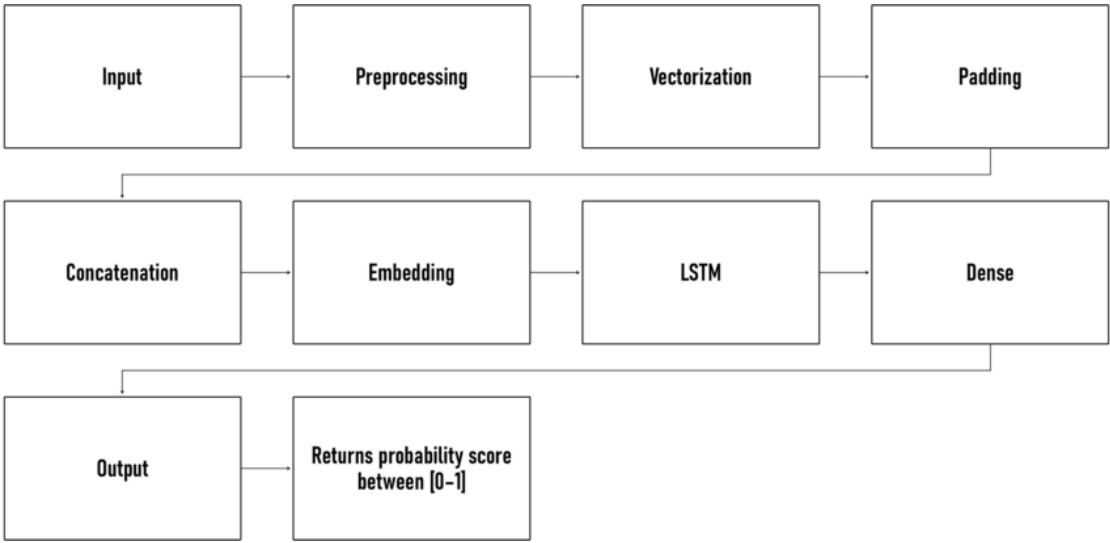


FIGURE 1: LSTM-RNN Training Flow
LSTM, Long Short-Term Memory; RNN, Recurrent Neural Networks

Results And Discussion

Experimental setup

All experiments were conducted on a system equipped with an AMD Ryzen 5 5600H processor, 16 GB RAM, and a dedicated NVIDIA RTX 3050 GPU for accelerated computing with the NVIDIA Compute Unified Device Architecture (CUDA) platform. The models were implemented using TensorFlow and trained for 100 epochs using an 80:20 train-validation split. The train-validation split took place after the augmentation process on the *model.fit* function of Keras' Sequential model with the *shuffle* parameter set to *True* to mitigate overfitting.

Training was performed using the Adam optimiser with a learning rate of 0.001 and a batch size of 32. Binary cross-entropy was used as the loss function because the final output of the model is between 0 and 1.

Results

Model performance was extensively evaluated using six key metrics: Accuracy, Binary Cross-Entropy Loss, Area Under the Curve (AUC), Precision, Recall, and the F1-Score. Figure 2 illustrates the trends for both training and validation accuracy across the investigated architectures. The distinct divergence observed in the non-augmented models differed from the stable, converging curves of the augmented models, underscoring the critical role of data augmentation in mitigating overfitting and enhancing model generalisation.

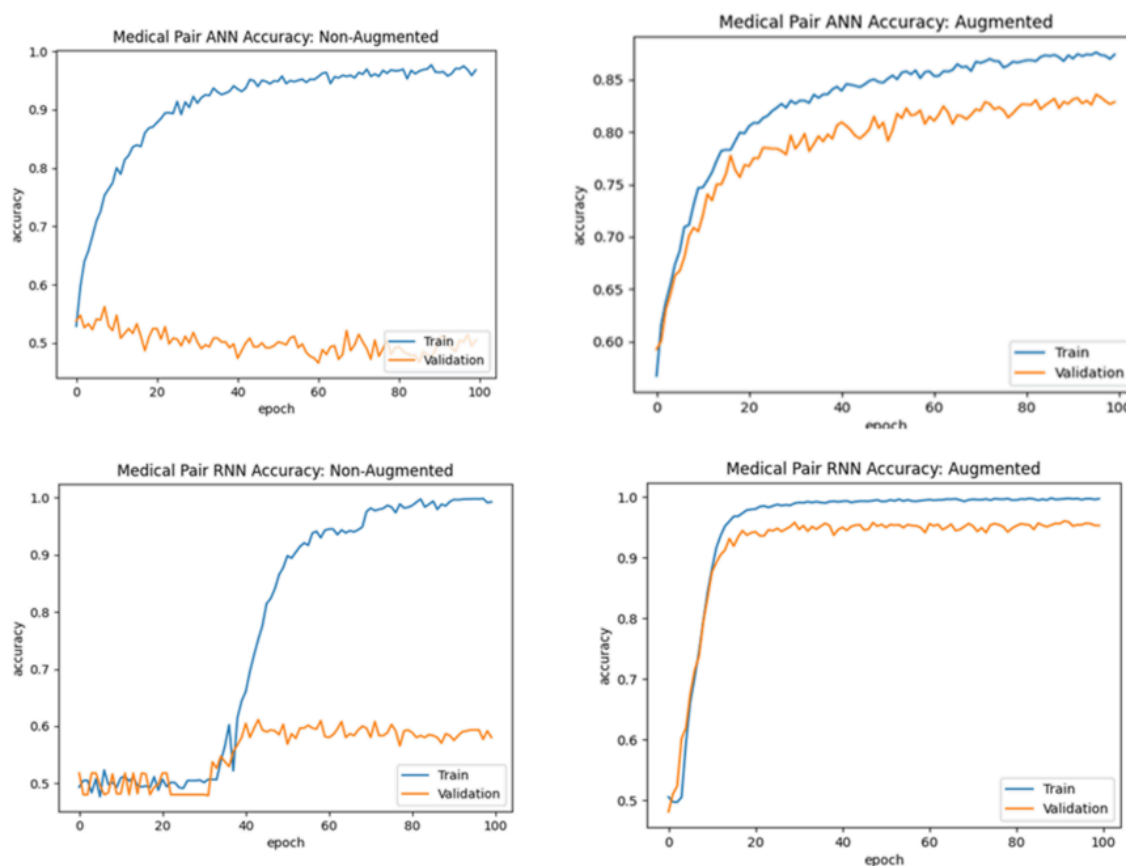


FIGURE 2: Training and Validation Accuracy Curve

Model performance summary

The augmented LSTM-RNN was the top performer, as shown in the performance comparison in Table 2. On validation data, it achieved 95.31% accuracy, 0.2931 loss, 97.39% AUC, 94.69% precision, 96.21% recall, and 95.44% F1-score. Without augmentation, validation accuracy plummeted to 58.03%, with an F1-score of 57.89%.

The augmented ANN achieved 82.87% validation accuracy, 0.4311 loss, 91.66% AUC, 84.98% precision, 79.72% recall, and 82.27% F1-score. The non-augmented ANN was the least effective, yielding a validation accuracy of only 50.49%.

Model	Train Acc	Train Loss	Train AUC	Val Acc	Val Loss	Val AUC
LSTM-RNN Augmented	99.70%	0.0115	99.95%	95.31%	0.2931	97.39%
LSTM-RNN Non-augmented	99.26%	0.0324	99.84%	58.03%	2.6612	59.96%
ANN Augmented	87.42%	0.2514	95.94%	82.87%	0.4311	91.66%
ANN Non-augmented	96.80%	0.0849	99.59%	50.49%	2.2138	49.13%

TABLE 2: Model Performance Comparison

ANN, Artificial Neural Network; AUC, Area Under the Curve; LSTM, Long Short-Term Memory; RNN, Recurrent Neural Network

Confusion matrix

To visually evaluate the efficiency of the different architectures, the models were applied to a test set of 500 samples. The predicted results were compared against the ground truth to generate confusion matrices using the *confusion_matrix* function within TensorFlow's *math* module. In Figure 3, the matrices were visualised as heatmaps using the Python Seaborn library to facilitate rapid interpretation. In these visualisations, a darker blue tint indicates a higher frequency of data points satisfying a specific condition. Notably, the "Medical Pair RNN Confusion Matrix: Augmented" demonstrates a superior concentration of true negatives (predicted 0 against actual 0) and true positives (predicted 1 against actual 1) when compared to the other matrices in the image.

How to cite this article:

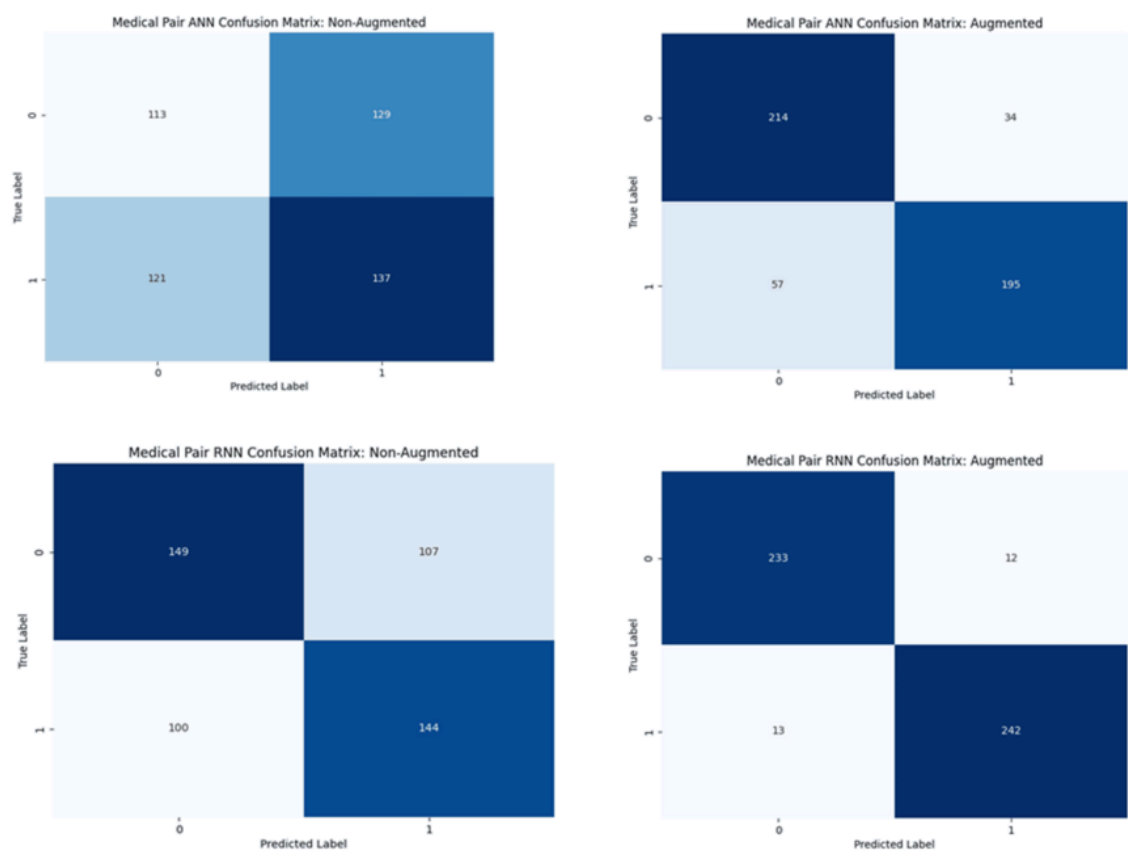


FIGURE 3: Confusion Matrices

Discussion

The choice to utilise a custom TextVectorization layer with a vocabulary size of 300 tokens rather than employing pre-trained word embeddings like GloVe or BERT was driven by the need to provide custom learning for the model within specific, evolving medical domains and localised linguistic contexts. For example, in certain regional contexts, phrases like "I am strong" are used as clinical affirmations of recovery rather than literal physical descriptions. By training the embedding layer directly on the provided dataset, the LSTM-RNN captures these specific nuances that generalised, Western-centric models might misinterpret as noise. Furthermore, while this study restricted lemmatisation to nouns to evaluate primary semantic anchors, the high performance suggests that the model effectively maps patient-reported intent through these core objects.

The significant performance gap between the LSTM-RNN and the ANN highlights the importance of the Embedding and LSTM layers. By processing text as a sequence, the LSTM-RNN was able to understand medical context more effectively [13]. The data also suggests that without augmentation, even sophisticated models fail to generalise due to limited dataset variety [10]. While successful, a limitation of this approach is the potential for semantic drift during random word substitution. Future work could benefit from transformer-based models like Bidirectional Encoder Representations from Transformers in the augmentation phase to preserve deeper contextual meanings [14].

How to cite this article:

Conclusions

This study confirms that LSTM-RNN architectures, when paired with robust data augmentation, can effectively predict semantic similarities in clinical text. By utilising embedding and LSTM layers, the model captured semantic nuances that traditional ANN structures missed. For this study, the implementation focused on the sequential processing of core nouns to capture medical context while maintaining computational efficiency. These findings offer a scalable solution for healthcare systems to manage unstructured data even when restricted by small, privacy-sensitive datasets. Future research may explore the inclusion of all parts of speech and broader alphanumeric data, such as dosages and vitals, to determine if additional linguistic categories further refine the model's predictive precision in more complex clinical settings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Daniel A. Folorunso, Adio T. Akinwale, Alaba O. Adejimi, Oluwaseun I. Adebisi

Acquisition, analysis, or interpretation of data: Daniel A. Folorunso

Drafting of the manuscript: Daniel A. Folorunso

Critical review of the manuscript for important intellectual content: Daniel A. Folorunso, Adio T. Akinwale, Alaba O. Adejimi, Oluwaseun I. Adebisi

Supervision: Adio T. Akinwale, Alaba O. Adejimi, Oluwaseun I. Adebisi

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data supporting this study are openly available in https://huggingface.co/datasets/curaihealth/medical_questions_pairs.

Acknowledgements

The dataset used in this study is publicly available from the Hugging Face repository: https://huggingface.co/datasets/curaihealth/medical_questions_pairs The data are anonymised and were accessed in accordance with the platform's data usage policies.

References

1. Singhal K, Azizi S, Tu T, et al.: [Large language models encode clinical knowledge](#). Nature. 2023, 620:172-180. [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)

How to cite this article:

Folorunso D A, Akinwale A T, Adejimi A O, et al. (April 28, 2026) A Semantic Text Similarity Model for Predicting Medical Transcription Data. Cureus J Comput Sci 3 : es44389-026-00062-6. DOI <https://doi.org/10.7759/s44389-026-00062-6>

2. Garcia AC, David GC, Chand D: [Understanding the work of medical transcriptionists in the production of medical records](#). Health Informatics Journal. 2010, 16:87-100. [10.1177/1460458210361936](#)
3. Callen J, McIntosh J, Li J: [Accuracy of medication documentation in hospital discharge summaries](#). International Journal of Medical Informatics. 2010, 79:58-64. [10.1016/j.ijmedinf.2009.09.002](#)
4. Zhang Y, Zerafa NM, Attard Montalto S: [Student and doctors' handwriting and transcription skills: how great is the potential for medical error?](#). Malta Medical Journal. 2020, 32:31-43.
5. Alonso V, Santos JV, Pinto M, et al.: [Health records as the basis of clinical coding: Is the quality adequate?](#). Health Information Management Journal. 2020, 49:28-37.
6. Feng Y: [Semantic textual similarity analysis of clinical text in the era of LLM](#). IEEE International Conference on Bioinformatics and Biomedicine. 2024, 1284-1289. [10.1109/CAI59869.2024.00227](#)
7. Dong X, Chowdhury S, Qian L, Li X, Guan Y, Yang J, Yu Q: [Deep learning for named entity recognition on Chinese electronic medical records](#). PLoS ONE. 2019, 14:e0216046. [10.1371/journal.pone.0216046](#)
8. Al-Askar H, Radi N, MacDermott Á: [Recurrent neural networks in medical data analysis and classifications](#). Applied Computing in Medicine and Health. Al-Jumeily D, Hussain A, Mallucci C, Oliver C (ed): Morgan Kaufmann, Waltham, MA; 2016. 147-165. [10.1016/B978-0-12-803468-2.00007-2](#)
9. Katte T: [Recurrent neural network and its various architecture types](#). International Journal of Research and Scientific Innovation (IJRSI). 2018, V:124-129.
10. Antunes R, Silva JF, Matos S: [Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings](#). SAC '20: Proceedings of the 35th Annual ACM Symposium on Applied Computing. 2020, 662-669. [10.1145/3341105.3373987](#)
11. McCreery C H, Katariya N, Kannan A, et al.: [Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs](#). KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020, 3458-3465.
12. Yu Y, Si X, Hu C, Zhang J: [A review of recurrent neural networks: LSTM cells and network architectures](#). Neural Computation. 2019, 31:1235-1270. [10.1162/neco_a_01199](#)
13. Soutner D, Müller L: [Application of LSTM neural networks in language modelling](#). Text, Speech, and Dialogue. Habernal I, Matoušek V (ed): Springer, Berlin, Heidelberg; 2013. 8082:105-112. [10.1007/978-3-642-40585-3_14](#)
14. Yang X, He X, Zhang H, Ma Y, Bian J, Wu Y: [Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models](#). JMIR Medical Informatics. 2020, 8:e19735. [10.2196/19735](#)

How to cite this article:

Folorunso D A, Akinwale A T, Adejimi A O, et al. (April 28, 2026) A Semantic Text Similarity Model for Predicting Medical Transcription Data. Cureus J Comput Sci 3 : es44389-026-00062-6. DOI <https://doi.org/10.7759/s44389-026-00062-6>