International Journal of Electrical and Computer Engineering (IJECE)
Vol. 15, No. 4, August 2025, pp. 4202~4212
ISSN: 2088-8708, DOI: 10.11591/ijece.v15i4.pp42021-4212

Enhancing multi-class text classification in biomedical literature by integrating sequential and contextual learning with BERT and LSTM

Oussama Ndama, Ismail Bensassi, Safae Ndama, El Mokhtar En-Naimi
DSAI2S Research Team, C3S Laboratory, Faculty of Sciences and Technologies of Tangier, Abdelmalek Essaâdi University, Tetouan, Morocco

Article Info

Article history:

Received Oct 20, 2024
Revised Mar 22, 2025
Accepted May 23, 2025

Keywords:

Bidirectional encoder representations from transformers
Biomedical text classification
Long short-term memory
Natural language processing
Sentence segmentation

ABSTRACT

Classification of sentences in biomedical abstracts into predefined categories is essential for enhancing readability and facilitating information retrieval in scientific literature. We propose a novel hybrid model that integrates bidirectional encoder representations from transformers (BERT) for contextual learning, long short-term memory (LSTM) for sequential processing, and sentence order information to classify sentences from biomedical abstracts. Utilizing the PubMed 200k randomized controlled trial (RCT) dataset, our model achieved an overall accuracy of 88.42%, demonstrating strong performance in identifying methods and results sections while maintaining balanced precision, recall, and F1-scores across all categories. This hybrid approach effectively captures both contextual and sequential patterns of biomedical text, offering a robust solution for improving the segmentation of scientific abstracts. The model's design promotes stability and generalization, making it an effective tool for automatic text classification and information retrieval in biomedical research. These results underscore the model's efficacy in handling overlapping categories and its significant contribution to advancing biomedical text analysis.

Corresponding Author:

Oussama Ndama
DSAI2S Research Team, C3S Laboratory, Faculty of Sciences and Technologies of Tangier, Abdelmalek Essaâdi University
Tetouan, Morocco
Email: oussama.ndama@etu.uae.ac.ma

1. INTRODUCTION

The surge in biomedical literature has made it increasingly difficult to extract valuable information efficiently from scientific papers. As the volume of biomedical research continues to grow, many abstracts are densely written, making them challenging to navigate and interpret [1], [2]. This issue is compounded by the absence of structured semantic frameworks in these abstracts, which hinders effective data retrieval and comprehension. While a variety of natural language processing (NLP) models have been developed to address these challenges [3]–[5], existing approaches often struggle to balance the demands of contextual understanding and sequential information within the text.

To address these challenges, this research introduces a novel hybrid model that integrates the strengths of bidirectional encoder representations from transformers (BERT) for contextual learning and long short-term memory (LSTM) for sequential learning. This combination aims to effectively segment and classify biomedical research paper abstracts, optimizing their readability while preserving essential information. The model leverages the robust contextual understanding capabilities of BERT and the sequential dependencies captured by LSTM, offering a more nuanced approach to text segmentation in biomedical literature.

This study builds upon the PubMed 200k RCT dataset, a widely used resource for sentence classification in medical abstracts [6]. While the original dataset comprises 2.3 million phrases from 200,000 randomized controlled trial abstracts, we have selected a subset of 500,000 sentences to train our model. Each sentence in this dataset is mapped to one of five predefined categories: background, objective, method, result, or conclusion, providing a structured framework for training and evaluation.

Our research leverages the computational capabilities of the Google Colab A100 GPU to efficiently train and fine-tune the hybrid model within real-world resource constraints. By harnessing both contextual and sequential features, this approach effectively enhances the readability and accessibility of scientific abstracts. Additionally, the method maintains computational efficiency, ensuring that the hybrid model remains viable for practical applications.

The primary objective of this work is to bridge gaps in current NLP approaches by offering a hybrid model that excels in understanding context while preserving the logical flow of biomedical text. This model improves the segmentation of biomedical abstracts by accurately dissecting their structural and semantic components. Furthermore, it aims to set a new standard for readability and comprehension within the field of biomedical research.

2. LITERATURE REVIEW

Biomedical text classification has seen significant advancements, particularly with the integration of machine learning and deep learning models. Rios and Kavuluru [7] (convolutional neural networks (CNNs) for biomedical text classification) demonstrated the effectiveness of CNNs in assigning medical subject headings (MeSH) to biomedical articles, outperforming traditional methods like logistic regression and support vector machines by improving macro F-scores. This work emphasized the advantages of CNNs in handling large feature spaces and complex biomedical text structures.

On the other hand, Dramé *et al.* [8] explored a k-nearest neighbors (kNN) based and explicit semantic analysis (ESA) based approach for large-scale biomedical text classification. Their kNN approach, combined with random forest (RF), achieved competitive performance with an F-measure of 0.55, while their ESA method underperformed. Their study highlighted the ongoing challenge of using partial information to classify documents in the biomedical domain.

In a broader review of biomedical text mining, Cohen [9] summarized the current progress in applying text mining techniques to tasks like named entity recognition, text classification, and hypothesis generation. They highlighted substantial advancements in computational methodologies and algorithms, enabling more effective extraction of meaningful patterns from biomedical texts. However, they noted that despite these advancements, considerable challenges persist, particularly in improving system usability for biomedical researchers and enhancing access to full-text articles, which are critical barriers limiting widespread adoption and practical utility of biomedical text mining tools.

Mondal introduced biomedical BERT-based adversarial example generation (BBAEG) [10], a novel adversarial example generation technique specifically for biomedical text classification. By leveraging BERT-masked language model (MLM) predictions and synonym replacement for biomedical entities, BBAEG demonstrated the potential of generating robust adversarial attacks that could expose vulnerabilities in current biomedical NLP models, highlighting the need for more resilient predictive systems. Further advancements in biomedical multi-label classification were explored by Zhang *et al.* [11], who introduced a multi-layer self-attention mechanism combined with BERT to enhance classification accuracy. Their model outperformed baselines in aspect category detection and biomedical document classification, showcasing the utility of self-attention for capturing complex dependencies in biomedical texts.

Neumann *et al.* [12] contributed significantly to the field with ScispaCy, a specialized Python library built upon spaCy, optimized specifically for processing biomedical texts. ScispaCy provides fast, scalable models achieving near-state-of-the-art performance across multiple biomedical NLP tasks, such as named entity recognition and parsing. Consequently, it serves as a robust and highly accessible tool, facilitating wider adoption among biomedical researchers and practitioners.

Document-level biomedical relation extraction was systematically addressed by Yuan *et al.* [13] through the introduction of the HTGRS framework, which employs hierarchical tree graphs and a dedicated relation segmentation module. Their framework strategically models interactions between entity pairs, significantly enhancing the accuracy of predicting relations across multiple biomedical entities. Experimental evaluations demonstrated that their method consistently outperformed previous state-of-the-art models, underscoring the value of structural modeling in biomedical relation extraction.

Duan *et al.* [14] tackled the challenge of sequential sentence classification in biomedical literature by proposing the boundary-aware dual biaffine model. Their innovative approach effectively leveraged document structural information, enabling precise detection of sentence boundaries and relationships. This method notably reduced classification errors, particularly in complex biomedical documents characterized by intricate sentence sequences and relationships.

Finally, Wang *et al.* [15] provided a comprehensive survey on the use of pre-trained language models (PLMs) in biomedical applications. They categorized the existing biomedical PLMs and discussed their applications in various tasks, noting both the advancements and limitations in the field. This survey emphasized the importance of cross-disciplinary collaboration to drive further innovation in biomedical NLP.

3. RESEARCH METHOD

This section outlines the methodology employed to develop and evaluate the hybrid model used for segmenting and classifying biomedical research paper abstracts. The model integrates BERT for contextual learning and LSTM for sequential learning to enhance the readability and segmentation of these abstracts. We utilized the PubMed 200k RCT dataset as a benchmark, focusing on a subset of 500,000 sentences to ensure computational efficiency while maintaining robust model performance. The model was trained using Google Colab's A100 GPU, adhering to constraints of resource availability and computational efficiency. In this section, we detail the dataset, model architecture, training procedure, and evaluation metrics used to assess the performance of the hybrid model.

3.1. Dataset

The dataset used in this study is the PubMed 200k RCT dataset, a large-scale resource designed for sequential sentence classification in biomedical abstracts. It comprises approximately 200,000 randomized controlled trial abstracts, totaling 2.3 million sentences. Each sentence is labeled with one of five predefined categories: background, objective, method, result, or conclusion. This dataset was released to address two key challenges: the lack of large-scale datasets for sequential short-text classification and the need for better tools to help researchers efficiently navigate lengthy biomedical abstracts.

For the purposes of this research, a subset of 500,000 sentences was sampled from the dataset to balance computational efficiency and model performance. Specifically, 22.61% of the original dataset was selected for training, which resulted in 500,102 samples in the training set and 29,493 samples in the test set. Each sample includes the following fields:

- Text: The sentence from the abstract.
- Chars: A character-level representation of the sentence.
- Order: The sequential position of the sentence within the abstract.
- Label: The sentence's category (one of the five predefined classes).

To preprocess the data, we employed a dual-level tokenization strategy, transforming the sentences at both the word and character levels. Character-level tokenization was achieved using a TextVectorization layer, configured with a custom vocabulary consisting of digits, punctuation marks, and ASCII characters. This approach captures the finer granularity of sentence structure, ensuring that every individual character contributes to the model's understanding of the input.

For word-level tokenization, the vocabulary was derived from a cleaned version of the dataset, where punctuation and unnecessary symbols were systematically removed to standardize the text [16], [17]. The resulting sequences of words were padded to align with the 95th percentile of sentence lengths, optimizing computational efficiency by setting a practical input length threshold. This careful approach ensured longer sentences were effectively accommodated without losing essential semantic information.

Label preprocessing involved converting categorical labels into numerical values through the use of a LabelEncoder, facilitating efficient computational handling. Subsequently, these numerical values were transformed via one-hot encoding, making them suitable for the multi-class classification task [18]. Consequently, this structured labeling approach enabled the model to classify each sentence accurately into one of the five predefined categories: background, objective, method, result, or conclusion.

The dataset was then split into training, validation, and test sets, with an 80/20 division between training and validation. This balanced split facilitated effective training and model tuning while preserving a portion of the data for unbiased evaluation. This preprocessing framework laid the foundation for the efficient training of the hybrid BERT-LSTM model, which is detailed in the following section.

3.2. Model architecture

The proposed model integrates both contextual and sequential learning to effectively classify sentences from biomedical abstracts. This hybrid architecture combines the strengths of BERT for contextual embedding and LSTM for sequential understanding. In addition, it incorporates sentence order information to further improve classification performance.

3.2.1. BERT encoder for contextual learning

The core of the model is the BERT encoder, a state-of-the-art model known for its ability to capture deep contextual relationships in text [19]. For this research, we use the pre-trained "*bert_base_en_uncased*" model, which is fine-tuned on the biomedical text dataset to adapt it to domain-specific language [20]. The BERT tokenizer and preprocessor are initialized with a sequence length of 256 tokens, ensuring that sentences are truncated or padded to a consistent length [21], [22]. The BERT encoder takes as input the tokenized sentence $x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$, where x_{ij} represents the j-th token in the i-th sentence. The BERT model generates a deep contextual embedding:

$$h_{\text{bert}} = \text{BERT}(x_i) \in R^{d_{\text{br}}} \tag{1}$$

where d_{bert} is the dimensionality of the BERT embedding space (768 in the base model). The pooled output from the BERT encoder, h_{bert}, is passed through a fully connected dense layer:

$$h_{\text{dense}} = \text{ReLU}(W_1 h_{\text{bert}} + b_1) \tag{2}$$

where $W_1 \in Rd$dne × $d$br and $b_1 \in Rd$dne are the learnable weights and biases of the dense layer. We apply L2 regularization to prevent overfitting, followed by a dropout layer to further improve generalization.

3.2.2. LSTM for sequential learning

In parallel with BERT's contextual embedding, the model incorporates an LSTM network to capture word-level sequential dependencies [23], [24]. The sentence x_i is first tokenized into words, which are then embedded into a 128-dimensional vector space using an embedding layer:

$$E_{\text{word}} = \text{Embedding}(x_i) \in Rn \times d\text{ebd} \tag{3}$$

where n is the number of words in the sentence and $d_{\text{embed}} = 128$ is the dimensionality of the word embedding space. The embedded sequence is processed by the LSTM layer to capture temporal dependencies:

$$h_{\text{lstm}} = \text{LSTM}(E_{\text{word}}) \in R^{d_{\text{lt}}} \tag{4}$$

where $d_{\text{lstm}} = 32$ represents the hidden state size of the LSTM. This captures sequential patterns that are not explicitly modeled by the BERT encoder. The LSTM output is passed through a fully connected dense layer with 16 units:

$$h_{\text{lstm\_dense}} = \text{ReLU}(W_2 h_{\text{lstm}} + b_2) \tag{5}$$

A dropout layer with a rate of 0.5 is applied to prevent overfitting during training. This component focuses on capturing sequential relationships between words [25], [26], providing a complementary view to BERT's contextual embeddings.

3.2.3. Incorporating sentence order information

The order in which sentences appear within an abstract is critical for understanding their role in the overall narrative. To leverage this, the sentence's position o_i is fed into a simple dense network:

$$h_{\text{order}} = \text{ReLU}(W_3 o_i + b_3) \tag{6}$$

where o_i is a scalar representing the sentence order, and $W_3 \in Rd$odr × 𝟙, $b3 \in Rd$odr are the weights and biases of the fully connected layer. This encoding provides additional insight into the function of the sentence based on its position in the abstract.

3.2.4. Fusion of features

The outputs from the BERT encoder, LSTM, and order processing layers are concatenated to create a unified representation of the sentence:

$$h_{\text{concat}} = \left[h_{\text{dense}}, h_{\text{lstm\_dense}}, h_{\text{order}}\right] \tag{7}$$

This combined vector $h_{\text{concat}} \in R^{d_{\text{cna}}}$ is then passed through a fully connected dense layer with 8 units and a ReLU activation function [27]:

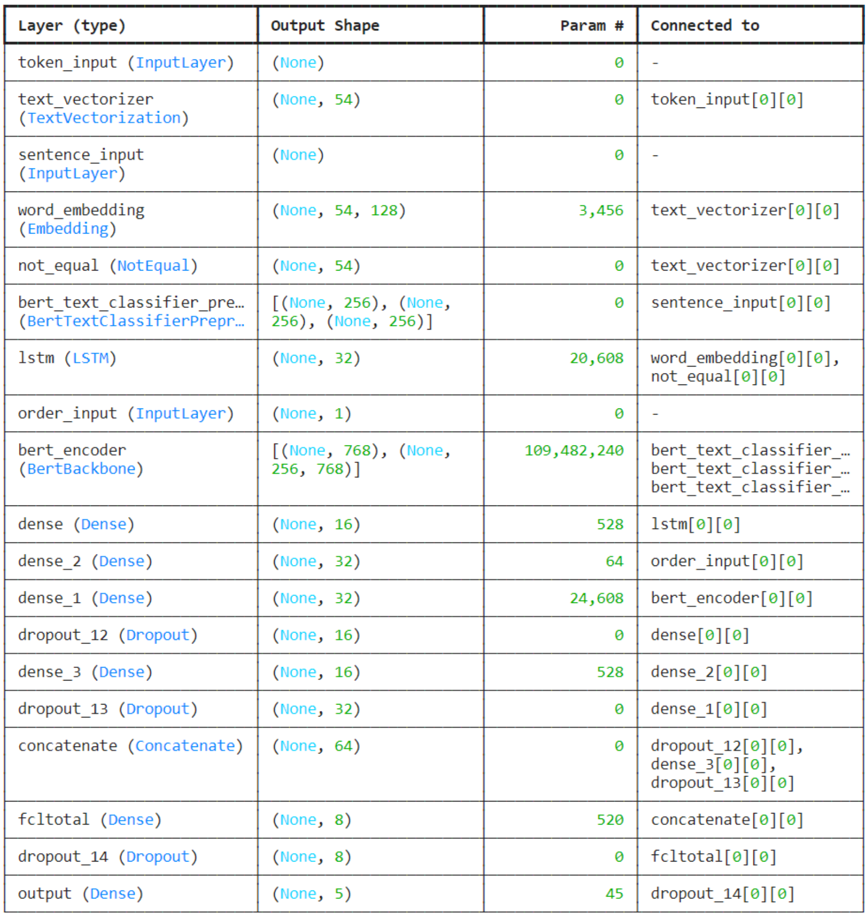
$$h_{\text{final}} = \text{ReLU}(W_4 h_{\text{concat}} + b_4) \tag{8}$$

To reduce overfitting, a dropout layer with a rate of 0.2 is added. Finally, a SoftMax layer [28] is applied to produce the probability distribution $p \in R^5$ over the five sentence categories:

$$p = \text{Softmax}(W_5 h_{\text{final}} + b_5) \tag{9}$$

where p_i represents the probability that the sentence belongs to category i.

Figure 1 provides a detailed breakdown of the model's architecture, outlining the various layers, their respective output shapes, and the number of parameters associated with each. Directly following the table, Figure 2 offers a visual representation of the model's structure. The architecture consists of three main components: the BERT encoder for capturing contextual information, the LSTM layer for modeling word-level sequences, and the dense layers processing sentence order information. These components are combined to form a comprehensive feature representation, which is then passed through fully connected layers for final classification. The model contains over 109 million trainable parameters, ensuring flexibility and capacity for learning complex patterns in biomedical text.



Layer (type)	Output Shape	Param #	Connected to
token_input (InputLayer)	(None)	0	-
text_vectorizer (TextVectorization)	(None, 54)	0	token_input[0][0]
sentence_input (InputLayer)	(None)	0	-
word_embedding (Embedding)	(None, 54, 128)	3,456	text_vectorizer[0][0]
not_equal (NotEqual)	(None, 54)	0	text_vectorizer[0][0]
bert_text_classifier_pre… (BertTextClassifierPrepr…	[(None, 256), (None, 256), (None, 256)]	0	sentence_input[0][0]
lstm (LSTM)	(None, 32)	20,608	word_embedding[0][0], not_equal[0][0]
order_input (InputLayer)	(None, 1)	0	-
bert_encoder (BertBackbone)	[(None, 768), (None, 256, 768)]	109,482,240	bert_text_classifier_… bert_text_classifier_… bert_text_classifier_…
dense (Dense)	(None, 16)	528	lstm[0][0]
dense_2 (Dense)	(None, 32)	64	order_input[0][0]
dense_1 (Dense)	(None, 32)	24,608	bert_encoder[0][0]
dropout_12 (Dropout)	(None, 16)	0	dense[0][0]
dense_3 (Dense)	(None, 16)	528	dense_2[0][0]
dropout_13 (Dropout)	(None, 32)	0	dense_1[0][0]
concatenate (Concatenate)	(None, 64)	0	dropout_12[0][0], dense_3[0][0], dropout_13[0][0]
fcltotal (Dense)	(None, 8)	520	concatenate[0][0]
dropout_14 (Dropout)	(None, 8)	0	fcltotal[0][0]
output (Dense)	(None, 5)	45	dropout_14[0][0]

Figure 1. Model summary and detailed breakdown of architecture

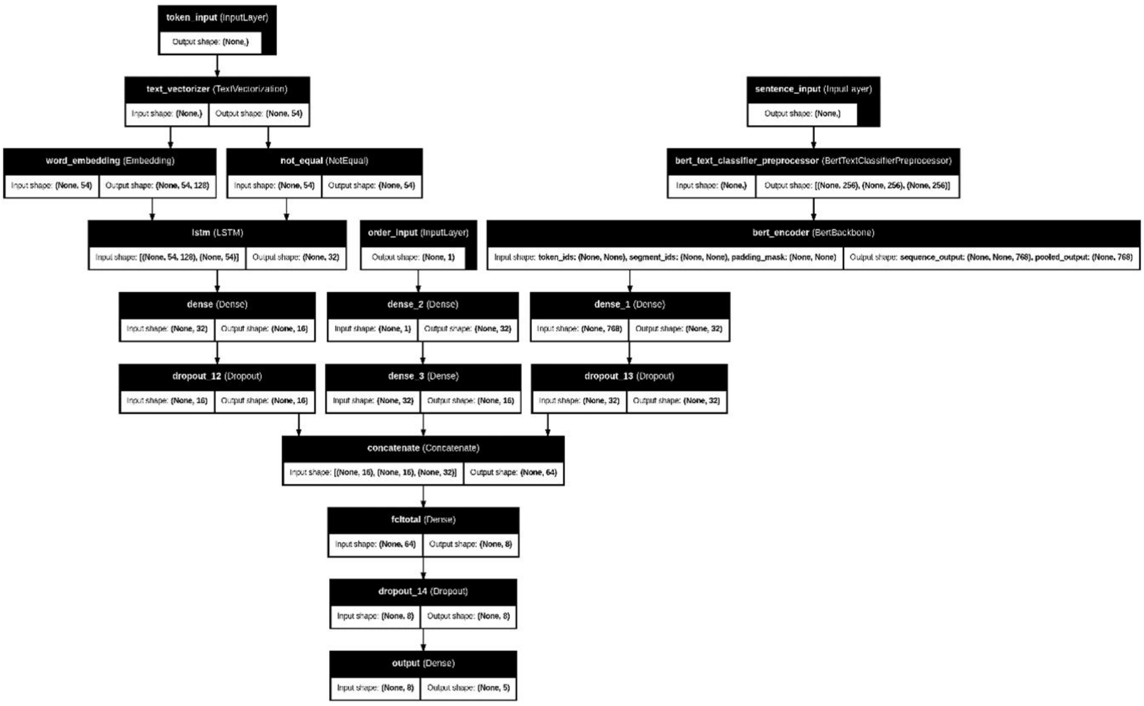


Figure 2. Architecture of the proposed hybrid model

3.3. Training procedure

To train the hybrid model, the dataset was split into training and validation sets, with 80% used for training and 20% for validation. The inputs included tokenized sentence text for both word-level and contextual-level embedding (using BERT) as well as sentence order information. These components were essential for feeding into the BERT encoder, LSTM, and order model layers, respectively.

3.3.1. Optimizer and loss function

The model was compiled using the Adam optimizer [29], with a learning rate of 2×10^{-5}. This small learning rate was selected to balance fast convergence and stable training. The model's performance was optimized using categorical crossentropy, the standard loss function for multi-class classification tasks [30]. This function is defined as:

$$L = -\sum_{i=1}^{5} y_i \log(p_i) \tag{10}$$

where y_i is the true label (one-hot encoded), and p_i is the predicted probability for each class.

3.3.2. Learning rate scheduler

To further optimize the learning process, a custom learning rate scheduler was used. For the first three epochs, the learning rate was kept constant, but after the third epoch, the learning rate was reduced by a factor of $e^{-0.1}$ at each epoch. This dynamic adjustment helped fine-tune the model as it approached convergence, slowing down learning to avoid overshooting optimal weights [31], [32]. The learning rate scheduler is defined as:

$$\text{lr\_scheduler}(\text{epoch,lr}) = \text{if epoch} < 3 \text{ then lr else lr} \cdot e^{-0.1} \tag{11}$$

This approach ensured that the model learned more aggressively in the initial epochs while gradually refining the weights as training progressed.

3.3.3. Callbacks

To ensure the model did not overfit and to speed up convergence, two key callbacks were employed:

- Early stopping: This callback monitored the validation accuracy and stopped training if there was no improvement for 3 consecutive epochs. The best weights were restored after training, ensuring that the model used the parameters from the epoch that yielded the highest validation accuracy.

- Learning rate scheduler: The custom learning rate scheduler described above dynamically adjusted the learning rate based on the training epoch.

3.3.4. Training configuration

The model was trained for 10 epochs, with a batch size of 16. This relatively small batch size allowed the model to effectively capture the complex relationships in the data without overwhelming memory. The training input consisted of:

- Sentence text: Passed to both the token and sentence models.
- Sentence order: Passed to the order model.

Training was performed on a Google Colab A100 GPU, leveraging the GPU's high computational power to speed up training and manage the resource-intensive nature of BERT fine-tuning.

3.4. Evaluation metrics

After training, the model was evaluated on the test set to assess its performance. The evaluation focused on measuring key metrics that demonstrate the model’s accuracy, precision, recall, and F1-score, providing a comprehensive view of its classification performance [33], [34]. This thorough assessment ensures the reliability and generalizability of the proposed model in handling various biomedical abstracts.

3.4.1. Test inputs

The test set consisted of 29,493 sentences, which were processed in the same way as the training and validation data. The model took as input:

- Sentence text: Tokenized and passed to both the BERT encoder and LSTM layers.
- Sentence order: A float value indicating the order of each sentence in the abstract.

3.4.2. Evaluation process and performance metrics

To evaluate the model, we build a custom function adapted to our hybrid model. This function computes the model's predictions by applying the SoftMax output to derive class probabilities and then selecting the class with the highest probability using argmax. These predictions were then compared against the true labels from the test set to compute the following metrics:

- Accuracy: The proportion of correct predictions out of all predictions. It provides a high-level measure of how well the model classified the sentences:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{12}$$

- Precision (Micro-Averaged): Precision measures how many of the predicted positive instances were correct, and is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}} \tag{13}$$

Micro-averaging computes this metric by aggregating contributions from all classes.

- Recall (Micro-Averaged): Recall measures how many of the actual positive instances were correctly identified by the model. It is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}} \tag{14}$$

- F1-score (Micro-Averaged): The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives:

$$\text{F1} = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{15}$$

4. RESULTS AND DISCUSSION

The performance of the proposed hybrid model was evaluated on the test set, and the results demonstrate a strong ability to classify sentences from biomedical abstracts into their respective categories. The evaluation metrics including accuracy, precision, recall, and F1-score indicate that the model effectively captures both contextual and sequential information in the text. The model achieved an overall accuracy of

88.42%, with a precision, recall, and F1-score of 88.42% across all categories, reflecting a balanced performance across different evaluation dimensions and highlighting the model's ability to correctly classify the sentences while maintaining a high degree of precision and recall.

The training and validation loss of the model showed consistent improvement over seven epochs, as illustrated in Figure 3. The training loss decreased from 1.14 to 0.41, while the validation loss followed a similar pattern, reducing from 0.50 to 0.39, with minor fluctuations toward the end. This convergence of both losses indicates that the model was effectively learning the patterns from the data without overfitting, maintaining stability throughout training. The total training time for this process was approximately 534.47 minutes (~8.9 hours), significantly reduced using a Google Colab A100 GPU, making the training efficient given the model's complexity and the size of the dataset.

As shown in Table 1, the model demonstrated strong performance in the 'Methods' and 'Conclusions' categories, with F1-scores of 93.56% and 88.79%, respectively, highlighting its ability to effectively differentiate these well-defined sections of biomedical abstracts. However, its performance in the 'Background' and 'Objective' categories was comparatively lower, with F1-scores of 74.03% and 70.47%, reflecting challenges in distinguishing these sections. The weighted averages for precision, recall, and F1-score 88.53%, 88.42%, and 88.45%, respectively underscore the model’s overall effectiveness in managing class imbalance within the dataset.

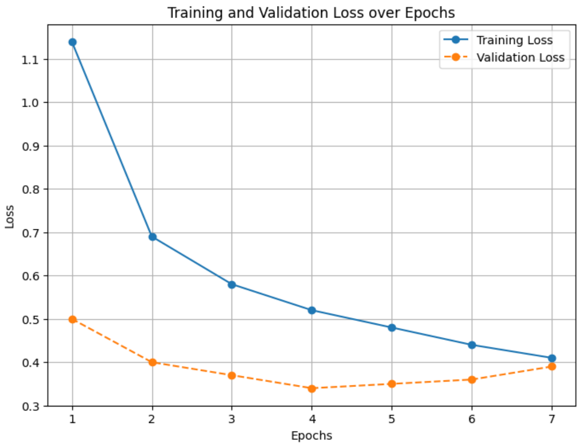


Figure 3. Training and validation loss over epochs

Table 1. Classification performance by category

	Accuracy	Recall	F1 Score	Support
Background	0.7221	0.7593	0.7402	2663
Conclusions	0.8661	0.9108	0.8879	4426
Methods	0.9361	0.9350	0.9356	9751
Objective	0.7134	0.6963	0.7047	2377
Results	0.9273	0.9005	0.9137	10276
Accuracy	-	-	0.8842	29493
Weighted Avg	0.8853	0.8842	0.8845	29493

The results indicate that the hybrid model, which combines BERT's contextual embeddings with LSTM's ability to capture sequential dependencies, is highly effective for multi-class sentence classification in biomedical abstracts. The high accuracy and balanced precision and recall across most categories demonstrate the model's robustness and generalization capability. The exceptional performance in the 'Methods' and 'Results' categories can be attributed to the distinct language and structure typically found in these sections, which the model could learn effectively.

The lower performance in the 'Background' and 'Objective' categories suggests that these sections may contain more nuanced language or share similarities with other sections, making them harder to classify. This overlap could be due to the introductory nature of these sections, where background information often sets the stage for the objectives of the study. Future work could focus on enhancing the model's ability to distinguish between these overlapping categories by incorporating additional linguistic features or leveraging domain-specific knowledge.

Moreover, the consistent decrease in training and validation loss over the epochs without significant fluctuations indicates that the model did not overfit and can generalize well to unseen data. This trend underscores the robustness of the training procedure and the stability of the model's architecture. The substantial training time further highlights the computational intensity of training such deep learning models; however, the use of high-performance computing resources like GPUs significantly alleviates this challenge.

In conclusion, the hybrid model demonstrates strong potential for automating the classification of sentences in biomedical literature, thereby facilitating more efficient information retrieval and knowledge extraction. Future enhancements could further improve the model's performance, particularly in challenging categories, making it an even more valuable tool for biomedical researchers and practitioners. Ongoing refinements to the model's architecture and optimization strategies may yield even more robust and scalable solutions in the long term.

5. CONCLUSION

In this study, we introduced a hybrid model that combines BERT for contextual learning, LSTM for sequential processing, and sentence order information to classify sentences in biomedical abstracts. The model achieved strong performance on the PubMed 200k RCT dataset, with an overall accuracy of 88.42% and balanced precision, recall, and F1-score across categories. It excelled in classifying methods and results sections, though further improvements could be made in distinguishing background and objective sentences. By effectively integrating both contextual and sequential information, our model demonstrates its potential for improving the readability and segmentation of complex biomedical texts. The use of fine-tuning, learning rate scheduling, and early stopping contributed to the model's convergence and stability, ensuring robust generalization to unseen data. This hybrid approach offers a valuable tool for enhancing the automatic processing of biomedical literature, enabling more efficient information retrieval for researchers. Future work could explore further refinements in classification accuracy, particularly in overlapping categories, and the integration of additional techniques such as attention mechanisms to enhance performance.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Oussama Ndama	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Ismail Bensassi						✓				✓				
Safae Ndama						✓				✓				
El Mokhtar En-Naimi		✓		✓		✓				✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the ACL Anthology at *https://www.aclweb.org/anthology/I17-2052.pdf*, reference: Dernoncourt and Lee (2017) [6]. PubMed 200k RCT: a Dataset for sequential sentence classification in medical abstracts. IJCNLP.

REFERENCES

[1] C. J. Cremin, S. Dash, and X. Huang, “Big data: Historic advances and emerging trends in biomedical research,” *Current Research in Biotechnology*, vol. 4, pp. 138–151, 2022, doi: 10.1016/j.crbiot.2022.02.004.

[2] Q. Jin *et al.*, “Biomedical question answering: A survey of approaches and challenges,” *ACM Computing Surveys*, vol. 55, no. 2, pp. 1–36, Feb. 2023, doi: 10.1145/3490238.

[3] T. August, L. L. Wang, J. Bragg, M. A. Hearst, A. Head, and K. Lo, “Paper Plain: Making medical research papers approachable to healthcare consumers with natural language processing,” *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 5, pp. 1–38, Oct. 2023, doi: 10.1145/3589955.

[4] G. Frisoni, G. Moro, and A. Carbonaro, “A survey on event extraction for natural language understanding: Riding the biomedical literature Wave,” *IEEE Access*, vol. 9, pp. 160721–160757, 2021, doi: 10.1109/ACCESS.2021.3130956.

[5] D. Demner-Fushman, N. Elhadad, and C. Friedman, “Natural language processing for health-related texts,” in *Biomedical Informatics*, Cham: Springer International Publishing, 2021, pp. 241–272.

[6] F. Dernoncourt and J. Y. Lee, “PubMed 200k RCT: A dataset for sequential sentence classification in medical abstracts,” *arXiv preprint arXiv:1710.06071*, 2017.

[7] A. Rios and R. Kavuluru, “Convolutional neural networks for biomedical text classification,” in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, Sep. 2015, pp. 258–267, doi: 10.1145/2808719.2808746.

[8] K. Dramé, F. Mougin, and G. Diallo, “Large scale biomedical texts classification: A kNN and an ESA-based approaches,” *Journal of Biomedical Semantics*, vol. 7, no. 1, p. 40, Dec. 2016, doi: 10.1186/s13326-016-0073-1.

[9] A. M. Cohen, “A survey of current work in biomedical text mining,” *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 57–71, Jan. 2005, doi: 10.1093/bib/6.1.57.

[10] I. Mondal, “BBAEG: Towards BERT-based biomedical adversarial example generation for text classification,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5378–5384, doi: 10.18653/v1/2021.naacl-main.423.

[11] X. Zhang, X. Song, A. Feng, and Z. Gao, “Multi-self-attention for aspect category detection and biomedical multilabel text classification with BERT,” *Mathematical Problems in Engineering*, vol. 2021, pp. 1–6, Nov. 2021, doi: 10.1155/2021/6658520.

[12] M. Neumann, D. King, I. Beltagy, and W. Ammar, “ScispaCy: Fast and robust models for biomedical natural language processing,” in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 319–327, doi: 10.18653/v1/W19-5034.

[13] J. Yuan, F. Zhang, Y. Qiu, H. Lin, and Y. Zhang, “Document-level biomedical relation extraction via hierarchical tree graph and relation segmentation module,” *Bioinformatics*, vol. 40, no. 7, Jul. 2024, doi: 10.1093/bioinformatics/btae418.

[14] J. Duan, H. Guo, H. Jiang, F. Guo, and J. Wang, “Boundary-aware dual biaffine model for sequential sentence classification in biomedical documents,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 21, no. 5, pp. 1202–1210, Sep. 2024, doi: 10.1109/TCBB.2024.3376566.

[15] B. Wang *et al.*, “Pre-trained language models in biomedical domain: A systematic survey,” *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–52, Mar. 2024, doi: 10.1145/3611651.

[16] S. J. Mielke *et al.*, “Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP,” *arXiv preprint arXiv:2112.10508*, 2021.

[17] B. Kim and H. J. Jang, “Word-level embedding to improve performance of representative Spatio-temporal document classification,” *Journal of Information Processing Systems*, vol. 19, no. 6, pp. 830–841, 2023, doi: 10.3745/JIPS.04.0296.

[18] N. Kosaraju, S. R. Sankepally, and K. Mallikharjuna Rao, “Categorical data: Need, encoding, selection of encoding method and its emergence in machine learning models—a practical review study on heart disease prediction dataset using Pearson correlation,” in *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 1*, 2023, pp. 369–382.

[19] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *Naacl-Hlt 2019*, no. Mlm, pp. 4171–4186, 2018, doi: 10.48550/arxiv.1810.04805.

[20] “BertTextClassifier model,” Keras, https://keras.io/api/keras_nlp/models/bert/bert_classifier/ (accessed Aug. 07, 2024).

[21] A. Merchant, E. Rahimtoroghi, E. Pavlick, and I. Tenney, “What happens to BERT embeddings during fine-tuning?,” *arXiv preprint arXiv:2004.14448*, pp. 33–44, 2020, doi: 10.18653/v1/2020.blackboxnlp-1.4.

[22] G. Puccetti, A. Miaschi, and F. Dell’Orletta, “How do BERT embeddings organize linguistic knowledge?,” in *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 2021, pp. 48–57, doi: 10.18653/v1/2021.deelio-1.6.

[23] D. Niu, Z. Xia, Y. Liu, T. Cai, T. Liu, and Y. Zhan, “ALSTM: Adaptive LSTM for durative sequential data,” in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov. 2018, pp. 151–157, doi: 10.1109/ICTAI.2018.00032.

[24] Yu Wang, “A new concept using LSTM neural networks for dynamic system identification,” in *2017 American Control Conference (ACC)*, May 2017, pp. 5324–5329, doi: 10.23919/ACC.2017.7963782.

[25] Y. Yu, X. Si, C. Hu, and J. Zhang, “A review of recurrent neural networks: LSTM cells and network architectures,” *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019, doi: 10.1162/neco_a_01199.

[26] G. Van Houdt, C. Mosquera, and G. Nápoles, “A review on the long short-term memory model,” *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5929–5955, Dec. 2020, doi: 10.1007/s10462-020-09838-1.

[27] A. M. Javid, S. Das, M. Skoglund, and S. Chatterjee, “A ReLU dense layer to improve the performance of neural networks,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 2810–2814, doi: 10.1109/ICASSP39728.2021.9414269.

[28] M. Wang, S. Lu, D. Zhu, J. Lin, and Z. Wang, “A high-speed and low-complexity architecture for SoftMax function in deep learning,” in *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Oct. 2018, pp. 223–226, doi: 10.1109/APCCAS.2018.8605654.

[29] S. Bock, J. Goppold, and M. Weiß, “An improvement of the convergence proof of the ADAM-optimizer,” *arXiv preprint arXiv:1804.10587*, 2018.

[30] J. Ghosh and S. Gupta, “ADAM optimizer and categorical Crossentropy loss function-based CNN method for diagnosing colorectal cancer,” in *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, Apr. 2023, pp. 470–474, doi: 10.1109/CISES58720.2023.10183491.

[31] C. Kim, S. Kim, J. Kim, D. Lee, and S. Kim, “Automated learning rate scheduler for large-batch training,” *arXiv preprint arXiv:2107.05855*, 2021.

[32] G. Ioannou, T. Tagaris, and A. Stafylopatis, “AdaLip: An adaptive learning rate method per layer for stochastic optimization,” *Neural Processing Letters*, vol. 55, no. 5, pp. 6311–6338, Oct. 2023, doi: 10.1007/s11063-022-11140-w.

[33] S. A. Hicks *et al.*, “On evaluation metrics for medical applications of artificial intelligence,” *Scientific Reports*, vol. 12, no. 1, p. 5979, Apr. 2022, doi: 10.1038/s41598-022-09954-8.

[34] Ž. Đ. Vujovic, “Classification model evaluation metrics,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120670.

BIOGRAPHIES OF AUTHORS

Oussama Ndama is a PhD student in DSAI2S (data science, artificial intelligence and smart systems research team), C3S Laboratory, Faculty of Sciences and Technologies (FST), Tangier, Morocco. He had his master’s in computer science and big data, Laureate of FST of Tangier. He is also a Business Intelligence Engineer with more than 5 years of experience in different multinational companies. The research topics of interest are smart systems, machine learning, deep learning, NLP, ANN, sentiment analysis, and smart cities. He can be contacted at email: oussama.ndama@etu.uae.ac.ma.

Ismail Bensassi is a PhD student in DSAI2S (data science, artificial intelligence and smart systems research team), C3S Laboratory, Faculty of Sciences and Technologies (FST), Tangier, Morocco. He is an engineer in computer science, Laureate of FST of Tangier. The research topics of interest are smart connection of user profiles in a big data context, multi-agent systems (MAS), case-based reasoning (CBR), ontology, machine learning, smart cities, and eLearning/MOOC/SPOC. He can be contacted at: bensassi.ismail@gmail.com.

Safae Ndama is a PhD student in DSAI2S (data science, artificial intelligence and smart systems research team), C3S Laboratory, Faculty of Sciences and Technologies (FST), Tangier, Morocco. She earned her master’s in computer science and big data from the FST of Tangier. With two years of experience as a data scientist, her research interests include smart systems, data science, artificial intelligence, sentiment analysis, and smart cities. She can be contacted via email: safae.ndama@etu.uae.ac.ma.

El Mokhtar En-Naimi is a full professor at the University of Abdelmalek Essaâdi (UAE), Faculty of Sciences and Technologies of Tangier (FSTT), Department of Computer Science. He was a temporary professor from 1999 to 2003 and a permanent professor from 2003/2004 until now. He served as Head of the Computer Science Department from October 2016 to December 2020 and was responsible for the Licence of Science and Technology in computer engineering (Licence LST-GI) from January 2012 to October 2016. He is the Chief of the Data Science, Artificial Intelligence and Smart Systems (DSAI2S) Research Team since the academic year 2022/2023. He is a founding member of both the LIST Laboratory (2008–2022) and the C3S Laboratory since the academic year 2022/2023 at the University of Abdelmalek Essaâdi, FST of Tangier, Morocco. He has been an expert evaluator with ANEAQ since 2016/2017 and an expert in academic programs at the Ministry of Higher Education and UAE University since 2012/2013. He is the author/co-author of several articles published in international journals in areas such as multi-agent systems (MAS), case-based reasoning (CBR), artificial intelligence (AI), machine learning (ML), deep learning (DL), big data, wireless sensor networks, smart cities, and more. He has supervised several doctoral theses and has served in various roles in international conferences. He is also an associate member of the ISCN – Institute of Complex Systems in Normandy, University of Le Havre, France, since 2009. He can be contacted at email: en-naimi@uae.ac.ma.