# Mental Health Sentiment Analysis: Exploring an Optimized BERT with Deep Encodings

**Abdul Rehman**
Cadet College Hasanabdal, Attock, Pakistan
69026@cch.edu.pk (corresponding author)

**Muzamil Ahmed**
Department of Computer Science, Namal University Mianwali, Pakistan
muzamil.ahmed@namal.edu.pk

**Hikmat Ullah Khan**
Department of Information Technology, University of Sargodha, Sargodha, Pakistan
dr.hikmat.niazi@gmail.com

**Amal Bukhari**
Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia
aabukhari@uj.edu.sa

**Ali Daud**
Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates
alimsdb@gmail.com

**Hussain Dawood**
School of Computing, Skyline University College, University City Sharjah, United Arab Emirates
dawood.hussain@skylineuniversity.ac.ae

*Received: 6 February 2025 | Revised: 19 February 2025 and 26 February 2025 | Accepted: 7 March 2025*

 *DOI: https://doi.org/10.48084/etasr.10469*

## ABSTRACT

**Deep learning technologies have significantly advanced the field of Natural Language Processing (NLP) in the recent past. A promising line of research in the existing study of text sentiment is the analysis of medical texts, which can potentially find numerous uses in medical diagnosis. However, sentiment analysis in the medical field remains challenging due to the complexity of medical language and context-aware interpretation of domain-specific ontologies. This paper focuses on the medical field and uses deep encoding techniques such as Fast-Text, Word2Vec, along with BERT and RoBERTa at the output layer for sentiment detection. The analysis specifically targets seven key mental health classes: Anxiety, Bipolar, Depression, Normal, Personality Disorder, Stress, and Suicidal. Text data were preprocessed using text cleaning, stop word elimination, and lemmatization to enhance the quality of the input data and improve the effectiveness of models. Experiments were carried out on a mental health dataset to analyze performance after the integration of deep encoding with diverse deep learning models. Based on the results, transformer-based models outperformed various other networks, achieving more than 95% accuracy. This study provides a basis for the selection of appropriate models in achieving accurate sentiment analysis within the medical field and is useful for research on designing efficient model frameworks.**

***Keywords-sentiment analysis; medical text; deep learning; natural language processing; transformer models***

## I. INTRODUCTION

Rapid growth in the use of social networks is transforming communication, especially within the healthcare setting [1]. This trend, among others, leads to oceans of data, which, if analysed through medical sentiment text mining using Natural Language Processing (NLP) and machine learning, can lead to understanding public sentiment regarding health issues [1, 2]. Sentiment analysis has been identified as an important avenue to explore in the extraction of meaningful insights from unstructured text data, such as tweets and blogs [3]. This tool has been considered to have great promise in understanding patient experiences, assessing health services, and improving general patient outcomes in the medical field [4]. However, the unique challenges posed by medical text, such as the use of specialized terminology and the need for high accuracy, have made it a complex and underexplored area [5, 6].

Sentiment analysis consists of tasks, each addressing a distinct level of text granularity and complexity. These include sentiment categorization at the document, phrase, and aspect levels, as well as multi-domain and multimodal approaches [7]. Document-level sentiment classification treats an entire document as one unit to identify the overall polarity of the sentiment, positive or negative. The challenge in this task is to model long texts to catch τηε semantic relations between two sentence pairs. Techniques such as the Hierarchical Attention Network (HAN) model and LSTM have been proposed to address these by focusing on the inter-document structure and inter-sentence relationships [8, 9]. Objective sentences provide information with no sentiment, while subjective sentences can provide polarity, either positively or negatively. Methods such as Weakly-supervised Deep Embedding (WDE) and Multi-level Sentiment-enriched Word Embedding (MSWE) have shown high classification accuracy based on CNN, LSTM, and MLP deep learning methods [10].

Aspect-level sentiment classification is the identification of sentiment-target pairs and their polarities on aspects or features of text. Attention mechanisms with LSTM models have been very effective in paying great attention to all parts of a sentence relative to various aspects [11]. Multidomain sentiment classification is the transfer of the classification knowledge developed in one domain to other domains. For example, the Domain Attention Model (DAM), using Bi-LSTM with attention mechanisms for domain-specific sentiments, has been proven rather effective in multi-domain data processing [12]. Multimodal sentiment classification combines several data modalities, such as text, audio, and visuals. Such methods are implemented through a weakly-supervised multi-modal deep learning model, combining CNNs and RNNs to achieve higher accuracy on heterogeneous data [13].

Depression, suicidal ideation, anxiety, stress, bipolar disorder, and personality disorders are medical conditions that can be extracted using sentiment analysis. These conditions affect how sentiments can be expressed and provide a lot of information for sentiment analysis. For example, when a person experiences depression or anxiety, his expression is typified by negative moods, hopelessness, or an increased state of tension and stress. On the other hand, a person with suicidal tendencies might manifest feelings of hopelessness and worthlessness, which would be very helpful for early prediction and prevention [14]. Sentiment analysis can also be used in bipolar disorder to detect changes from manic to depressive episodes in content and tone in patients' communication. Personality disorders are most commonly manifested as persistent patterns of thought and behavior that can be recognized and tracked through sentiment analysis. This not only demonstrates the power to develop knowledge in patient experiences, but also supports the early diagnosis and management of such mental health states [12, 15].

Earlier methods, such as polarity classification, lexicon-based, and emotion classification, are building blocks. However, more advanced methods are devised to tackle certain difficulties within medical texts [16]. The study in [17] expanded on these methods using supervised machine techniques for hand-crafted relevant features to classify polarity after filtering the sentiment. A significant contribution in this direction is the VADER (Valence Aware Dictionary for Sentiment Reasoning) lexicon, which has now become a standard tool in the field of sentiment analysis [18]. A comparative study of sentiment analysis techniques in [19] further underlined the strength of SVM in this aspect. Another important area of focus within medical sentiment analysis has been on emotion classification. In [20], emotions were classified as hope, worry, or gratitude using a Naive Bayes classifier with features computed from the WordNet-Affect lexicon. In [21], ADR mentions were extracted from user narratives on Twitter and PatientsLikeMe to study two methods: the use of SIDER/MedDRA ADR lexicons and system condition terms developed by a Medicomp, Inc. system with maintained databases. Traditional machine learning approaches have focused on detectable ADRs through the use of NLP techniques and pre-existing lexicons [22]. However, deep learning has opened up a new paradigm that can improve the accuracy of sentiment classification [23].

## II. PROPOSED METHOD

Figure 1 represents the framework of the proposed method, which consists of four main components, including text preprocessing, feature engineering, proposed models, and performance evaluation.

### A. Text Preprocessing

This step aims to refine the text by correcting errors, inconsistencies, and noise within the data. First, duplicate entries were identified and removed to overcome bias during the analysis. Second, spelling errors and typos were corrected within the data to avoid misleading inferences. The next step involved stop-word removal. Common words, such as "and, the, is, and in," are known as stop-words, which are very frequent but carry little meaning and can clutter an analysis. This reduces dimensionality and makes efficiency and relevance possible in the results. Lemmatization replaced the actual words with their meaningful root words. Such a standardization can help recognize other variations of the same word and thus improve the consistency and accuracy of sentiment analysis, especially in medical texts.

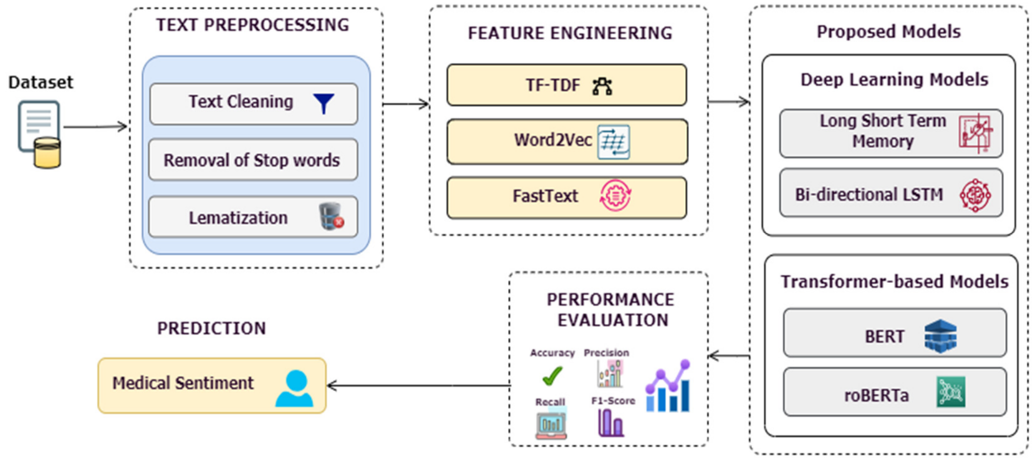


Fig. 1. Overview of the proposed method for medical sentiment analysis.

## *B. Feature Engineering*

### *1) TF-IDF*

This is an approach for generating vector representations of text by assessing the relevance of a word inside a document in comparison to a collection of documents. The TF-IDF score is calculated by multiplying two components: Term Frequency (TF) and Inverse Document Frequency (IDF). TF is determined using:

$$TF(t,d) = \frac{f(t,d)}{N_d} \tag{1}$$

where $f(t,d)$ is the frequency of the term $t$ in document $d$, and $N_d$ is the total number of terms in that document. IDF measures the rarity of a term across the corpus, calculated using:

$$IDF(t,D) = \log\left(\frac{N}{f(t,D)}\right) \tag{2}$$

where $N$ is the total number of documents in the corpus, and $f(t,D)$ is the number of documents containing the term $t$. The final TF-IDF score for a term in a document is calculated by multiplying its TF and IDF values:

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D) \tag{3}$$

### *2) Word2Vec*

This encoding catches semantic relationships between the words, allowing similar words to possess close vector representations. For a given target word $w_i$ and its context $C = \{w_{i-K}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+K}\}$, the model estimates the probability of each context word $w_j$ given the target word $w_i$. This probability is calculated using:

$$P(w_j|w_i) = \frac{\exp(v_{w_j}^{\top} v_{w_i})}{\sum_{w \in V} \exp(v_w^{\top} v_{w_i})} \tag{4}$$

where $v_{w_i}$ denotes the vector representation of the target word $w_i$, $v_{w_j}$ is the vector representation of the context word $w_j$, and $V$ represents the vocabulary. The objective is to maximize this probability for all context words given a target word. The loss function, based on the negative log-likelihood, is:

$$Loss = -\sum_{w_j \in C} v_{w_j}^{\top} v_{w_i} + \log\left(\sum_{w \in V}\left(v_w^{\top} v_{w_i}\right)\right) \tag{5}$$

The loss function penalizes the difference between the predicted probabilities of context words and their actual occurrences, leading to an update in the word vectors. Representation of words as dense continuous vectors gives Word2Vec better performance in sentiment analysis, even in domain-specific discourse, such as medical text, with complicated and diverse terminologies [17].

### *3) Fast-Text*

This model represents a word as the sum of character $n$-grams and not atomically such as Word2Vec. This allows the model to generate representations from constituent parts of words, making it very efficient in handling morphologically rich language. For a given target word $w_i$ and its context $C = \{w_{i-K}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+K}\}$, the model estimates the probability of each context word $w_j$ given the target word $w_i$ using:

$$P(\, w_j \mid w_i \,) = \frac{\exp\left(v_w^{\top} v_{w_i}\right)}{\sum_{w \in V} \exp(v_w^{\top} v_{w_i})} \tag{6}$$

It maximizes this probability for all context words, given any target word, to provide high-quality word embeddings.

## *C. Deep Learning Model*

### *1) LSTM*

LSTM can capture both short-term and long-term dependencies in temporal data. Each LSTM consists of interconnected memory blocks, which include three key units: the forget gate, the input gate, and the output gate [13]. Long-term memory, or the cell state, is governed by the forget gate $f(t)$, which determines which parts of the long-term state to discard. The input gate $i(t)$ manages the incorporation of new information into the cell state, while the output gate $o(t)$ regulates the generation of the current state output. The outputs of these gates are calculated using:

$$f(t) = \varphi\left(w_{fx} \cdot x(t) + w_{fh} \cdot h(t-1) + b_f\right) \tag{7}$$

$$(t) = \varphi\left(w_{ix} \cdot x(t) + w_{ih} \cdot h(t-1) + b_i\right) \tag{8}$$

$$o(t) = \varphi\,(w_{ox} \cdot x(t) + w_{oh} \cdot h(t-1) + b_o) \quad (9)$$

where $\varphi(\cdot)$ represents a nonlinear activation function, typically the sigmoid function. Alongside the input gate, the candidate state $\tilde{C}(t)$ is calculated as:

$$\tilde{C}(t) = \tanh(w_{cx} \cdot x(t) + w_{ch} \cdot h(t-1) + b_c) \quad (10)$$

$$C(t) = f(t) \odot C(t-1) + i(t) \odot \tilde{C}(t) \quad (11)$$

$$h(t) = o(t) \odot \tanh(\mathrm{C(t)}) \quad (12)$$

where $\tanh(\cdot)$ is the hyperbolic tangent activation function and $\odot$ denotes the point multiplication operation between two vectors. The terms $w_{fx}$, $w_{ix}$, $w_{ox}$, and $w_{cx}$ represent the input weights for each gate, while $w_{fh}$, $w_{ih}$, $w_{oh}$, and $w_{ch}$ are the recurrent weights, and $b_f$, $b_i$, $b_o$, and $b_c$ are the respective biases [15].

*2) Bi-LSTM*

A Bi-LSTM network uses two LSTM layers: one processes the input sequence in the forward direction and the other processes it in the backward direction. Figure 2 represents the basic structure of a Bi-LSTM network.

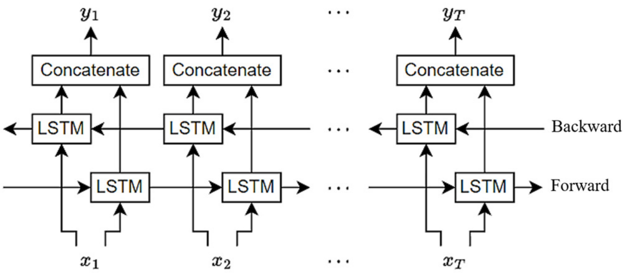


Fig. 2. Structure of BiLSTM for medical sentiment analysis.

For a given sequence $x = (x_1, x_2, \ldots, x_T)$, the forward LSTM computes a sequence of hidden states $\overrightarrow{h_t}$, while the backward LSTM computes $\overleftarrow{h_t}$. At each time step $t$, the final output $h_t$ is a concatenation of these two hidden states, represented mathematically as $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. Here, $\overrightarrow{h_t}$ captures information from the start up to $t$, while $\overleftarrow{h_t}$ captures information from the end up to $t$.

*D. Transformer-based Model*

This study utilized transformer-based methods, including BERT and RoBERTa, for empirical analysis of medical sentiment.

*1) BERT*

This study employs a masked language modeling approach as shown in Figure 3, where it predicts missing words by masking 15% of the words in the input text. Each transformer block receives a phrase with a masked word and aims to predict that missing word. BERT is bidirectional, meaning that it learns to predict the missing word by considering both the preceding and following words in the context. This allows it to effectively capture both left (previous) and right (subsequent) contexts for each word. The embeddings produced by BERT are context-aware, as the missing word is inferred based on its surrounding words.

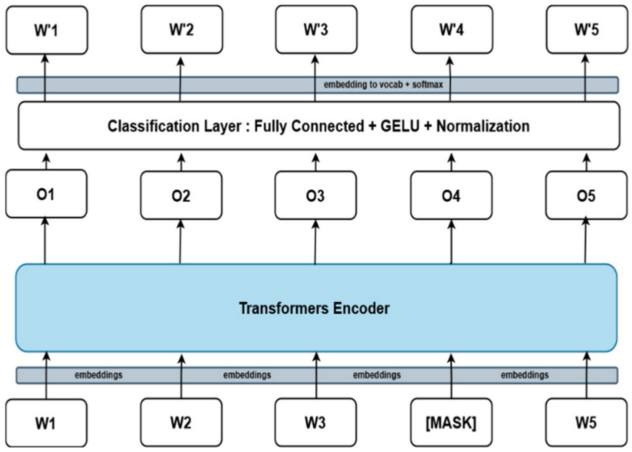


Fig. 3. Structure of BERT for medical sentiment classification.

In addition to generating embeddings for individual words, BERT also produces an embedding for the entire sentence. Typically, BERT is not trained from scratch but is pre-trained on extensive internet data. It uses a multilayer bidirectional transformer encoder to encode the input text into a high-level semantic space.

*2) Robustly Optimized BERT (RoBERTa)*

RoBERTa solely focuses on the Masked Language Modeling (MLM) objective, removing the next sentence prediction task. MLM involves randomly masking 15% of input tokens and training the model to predict them using contextual information, using dynamic masking that changes patterns each epoch. The MLM objective is to maximize the probability of the masked tokens given their context, calculated as:

$$\mathcal{L}_{MLM} = \sum_{i=1}^{k} \log P(m_i | \mathbf{X}_{\{m_i\}}) \quad (13)$$

RoBERTa's optimizations include training on a 160GB dataset, ten times larger than BERT's, and extending training duration with larger batch sizes. It employs the same transformer architecture, where each self-attention layer calculates attention scores using scaled dot-product attention:

$$A = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (14)$$

This architecture captures bidirectional context, enhancing dependency recognition.

## III. RESULTS

*A. Experimental Setup*

*1) Dataset*

This study used a publicly available dataset, called Sentiment Analysis for Mental Health [24], to evaluate the proposed model for medical sentiment analysis. It is a comprehensive collection of 51,074 entries meticulously curated to analyze various mental health conditions through user-generated statements. This dataset integrates data from multiple standard corpora, including the 3k conversations dataset for chatbot, depression reddit cleaned, human stress prediction, predicting anxiety in mental health data, mental health dataset bipolar, reddit mental health data, students anxiety and depression dataset, suicidal mental health dataset, and suicidal tweet detection dataset. Each entry is tagged with one of seven distinct mental health statuses: Normal, Depression, Suicidal, Anxiety, Stress, Bi-Polar, and Personality Disorder.

*2) Evaluation Measures*

Accuracy, precision, recall, and F1 score were used to evaluate model performance. Accuracy is defined as the ratio of correctly predicted instances to the total instances in the dataset:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (15)$$

where TP and TN refer to the number of correctly predicted positive and negative instances, and FP and FN refer to the number of incorrectly predicted positive and negative instances, respectively. Precision is defined as the ratio of TP predictions to the total predicted positives, which includes both TP and FP:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (16)$$

Recall is defined as the ratio of TP predictions to the total actual positives, which includes both TP and FN:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (17)$$

The F1 score is the harmonic mean of precision and recall, taking both FP and FN into account:

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (19)$$

*B. Results of Deep Learning Models*

This section presents results of LSTM and Bi-LSTM with TF-IDF, Word2Vec, and Fast-Text for medical sentiment analysis. According to Table I, LSTM achieved 67.12%, 74.75% and 70.80% accuracy over 30 epochs with TF-IDF, Word2Vec, and Fast-Text embeddings, respectively.

Similarly, as shown in Table II, the BiLSTM model showed an increasing trend, with an accuracy of 70.50%, 78.42% and 81.08% with TF-IDF, Word2Vecm, and FastText, respectively. The best performance of the BiLSTM was obtained using FastText embeddings at the end of 20 epochs.

BiLSTM, due to bidirectional learning, outperformed LSTM across all embedding methods. The significant performance of FastText embeddings highlights its ability to generate sub-word-level embeddings, allowing the model to better understand out-of-vocabulary terms and capture more fine-grained word semantics.

TABLE I. LSTM RESULTS ON MULTICLASS MEDICAL SENTIMENT ANALYSIS

| Embedding | Epochs | Accuracy % | Precision % | Recall % | F1-score % |
|---|---|---|---|---|---|
| TF-IDF | 5 | 62.34 | 63.12 | 61.80 | 62.45 |
| | 10 | 64.89 | 65.50 | 64.20 | 64.84 |
| | 15 | 66.02 | 66.80 | 65.90 | 66.35 |
| | 20 | 68.58 | 68.90 | 68.40 | 68.65 |
| | 25 | 67.50 | 68.10 | 67.00 | 67.55 |
| | 30 | 67.12 | 67.70 | 66.50 | 67.10 |
| Word2Vec | 5 | 71.25 | 71.90 | 70.80 | 71.35 |
| | 10 | 72.91 | 73.40 | 72.60 | 73.00 |
| | 15 | 74.12 | 74.60 | 73.80 | 74.20 |
| | 20 | 76.24 | 76.80 | 75.90 | 76.35 |
| | 25 | 75.50 | 76.00 | 75.20 | 75.60 |
| | 30 | 74.75 | 75.30 | 74.40 | 74.85 |
| FastText | 5 | 65.35 | 66.00 | 64.90 | 65.45 |
| | 10 | 68.20 | 68.70 | 67.90 | 68.30 |
| | 15 | 71.10 | 71.60 | 70.60 | 71.10 |
| | 20 | 72.91 | 73.40 | 72.60 | 73.00 |
| | 25 | 71.45 | 72.00 | 71.10 | 71.55 |
| | 30 | 70.80 | 71.30 | 70.20 | 70.75 |

TABLE II. I-LSTM RESULTS ON MULTICLASS MEDICAL SENTIMENT ANALYSIS

| Embedding | Epochs | Accuracy % | Precision % | Recall % | F1-score % |
|---|---|---|---|---|---|
| TF-IDF | 5 | 63.45 | 64.10 | 62.90 | 63.50 |
| | 10 | 66.78 | 67.40 | 66.10 | 66.75 |
| | 15 | 68.90 | 69.50 | 68.30 | 68.90 |
| | 20 | 71.39 | 71.90 | 70.90 | 71.40 |
| | 25 | 70.50 | 71.00 | 69.90 | 70.45 |
| | 30 | 69.25 | 69.80 | 68.70 | 69.25 |
| Word2Vec | 5 | 70.12 | 70.70 | 69.60 | 70.15 |
| | 10 | 73.56 | 74.20 | 73.10 | 73.65 |
| | 15 | 75.82 | 76.40 | 75.20 | 75.80 |
| | 20 | 78.42 | 79.00 | 78.10 | 78.55 |
| | 25 | 77.10 | 77.60 | 76.50 | 77.05 |
| | 30 | 76.35 | 76.90 | 75.80 | 76.35 |
| FastText | 5 | 73.80 | 74.40 | 73.20 | 73.80 |
| | 10 | 77.90 | 78.50 | 77.40 | 77.90 |
| | 15 | 79.55 | 80.10 | 79.00 | 79.55 |
| | 20 | 81.08 | 81.60 | 80.50 | 81.05 |
| | 25 | 80.25 | 80.80 | 79.70 | 80.25 |
| | 30 | 78.90 | 79.50 | 78.30 | 78.85 |

*C. Results of Transformer-Based Models*

Tables III and IV show the results of BERT and RoBERTa, respectively, over diverse embeddings. BERT achieved 81.26%, 95.71% and 91.05% accuracy with TF-IDF, Word2Vec, and FastText, respectively. BERT showed good performance with Word2Vec on 25 epochs. On another hand, RoBERTa achieved 87.62%, 92.08%, and 94.47% accuracy with TF-IDF, Word2Vec, and FastText, respectively. RoBERTa also performed well with FastText at 20 epochs.

Figures 4 and 5 represent the confusion matrices of BERT and RoBERTa for the detection of medical sentiment. In the given confusion matrices, the correct predictions are shown in the diagonal, and off-diagonal elements show misclassifications. Figures 7 and 8 compare deep learning and transformer-based models in terms of accuracy and F1-score, respectively.

TABLE III. BERT RESULTS ON MULTICLASS MEDICAL SENTIMENT ANALYSIS

| Embedding | Epochs | Accuracy % | Precision % | Recall % | F1-score % |
|---|---|---|---|---|---|
| TF-IDF | 5 | 65.32 | 65.90 | 64.70 | 65.30 |
| | 10 | 69.45 | 70.00 | 68.90 | 69.40 |
| | 15 | 74.70 | 75.30 | 73.90 | 74.60 |
| | 20 | 78.85 | 79.40 | 78.20 | 78.80 |
| | 25 | 81.26 | 81.80 | 80.90 | 81.35 |
| | 30 | 80.10 | 80.60 | 79.50 | 80.05 |
| Word2Vec | 5 | 82.40 | 83.00 | 81.80 | 82.40 |
| | 10 | 88.50 | 89.10 | 88.00 | 88.55 |
| | 15 | 92.35 | 92.90 | 91.80 | 92.35 |
| | 20 | 95.71 | 96.20 | 95.40 | 95.80 |
| | 25 | 94.30 | 94.80 | 93.90 | 94.35 |
| | 30 | 93.10 | 93.60 | 92.70 | 93.15 |
| FastText | 5 | 76.20 | 76.80 | 75.60 | 76.20 |
| | 10 | 84.40 | 85.00 | 83.80 | 84.35 |
| | 15 | 89.55 | 90.10 | 88.90 | 89.50 |
| | 20 | 91.05 | 91.50 | 90.60 | 91.05 |
| | 25 | 90.30 | 90.80 | 89.90 | 90.35 |
| | 30 | 88.75 | 89.30 | 88.10 | 88.70 |

TABLE IV. ROBERTA RESULTS ON MULTICLASS MEDICAL SENTIMENT ANALYSIS

| Embedding | Epochs | Accuracy % | Precision % | Recall % | F1-score % |
|---|---|---|---|---|---|
| TF-IDF | 5 | 71.85 | 72.50 | 71.10 | 71.80 |
| | 10 | 75.30 | 75.90 | 74.70 | 75.30 |
| | 15 | 80.45 | 81.10 | 79.80 | 80.45 |
| | 20 | 84.70 | 85.20 | 84.20 | 84.70 |
| | 25 | 87.62 | 88.10 | 87.20 | 87.65 |
| | 30 | 86.30 | 86.80 | 85.70 | 86.25 |
| Word2Vec | 5 | 78.20 | 78.80 | 77.50 | 78.15 |
| | 10 | 83.40 | 84.00 | 82.70 | 83.35 |
| | 15 | 88.55 | 89.10 | 88.00 | 88.55 |
| | 20 | 92.08 | 92.60 | 91.80 | 92.20 |
| | 25 | 91.30 | 91.80 | 90.90 | 91.35 |
| | 30 | 90.10 | 90.60 | 89.70 | 90.15 |
| FastText | 5 | 83.45 | 84.00 | 82.90 | 83.45 |
| | 10 | 88.90 | 89.50 | 88.20 | 88.85 |
| | 15 | 92.60 | 93.10 | 92.00 | 92.55 |
| | 20 | 94.47 | 94.90 | 94.00 | 94.45 |
| | 25 | 93.80 | 94.30 | 93.40 | 93.85 |
| | 30 | 92.20 | 92.70 | 91.80 | 92.25 |

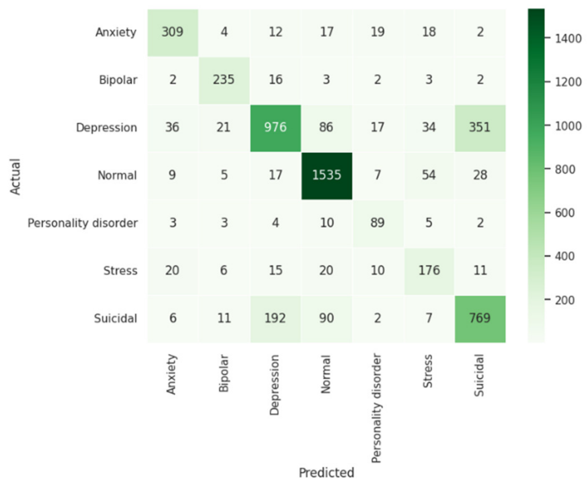


Fig. 4. Confusion matrix of BERT for medical sentiment analysis.

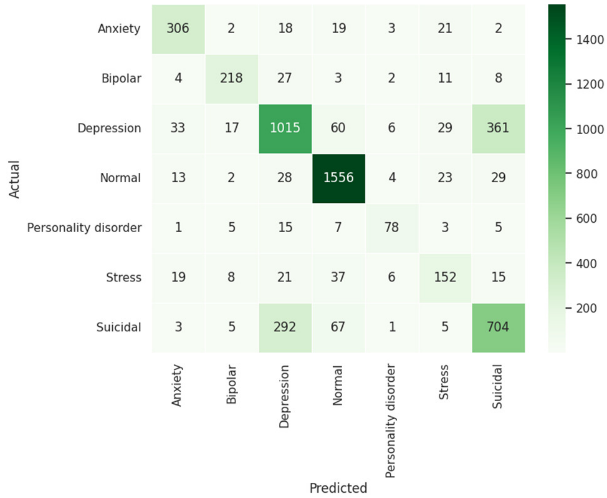


Fig. 5. Confusion matrix of RoBERTa for medical sentiment analysis.

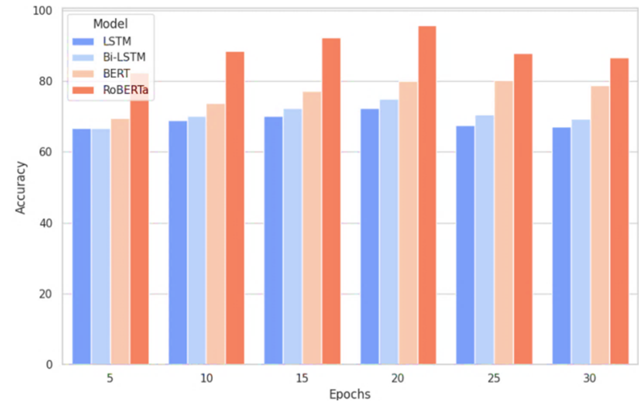


Fig. 6. Accuracy of deep learning and transformer-based models over different numbers of epochs.

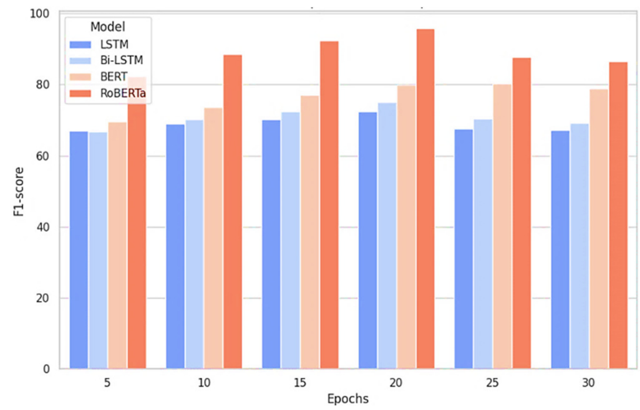


Fig. 7. F1-score of deep learning and transformer-based models over different numbers of epochs.

RoBERTa performed better than all other models in terms of accuracy and F1-score with most embeddings. However, both transformers highlight a significant increase in accuracy compared to the deep learning models. This can be explained by the transformers' self-attention mechanism, which allows the model to focus on key parts of the input while making predictions, thus capturing more contextual information.

## IV. CONCLUSION

This study aimed to advance sentiment analysis in the medical domain by applying state-of-the-art deep learning methods to accurately interpret complex medical text. Extensive preprocessing ensured that the data was cleaned and ready for analysis. The study employed diverse embedding techniques, such as Word2Vec and FastText, to learn the semantics of medical texts. A wide array of models was examined, from LSTM and BiLSTM to more advanced transformer-based models, such as BERT and RoBERTa. The results of the LSTM and BiLSTM models were on par with expectations and effectively served as a baseline. The transformer-based models outperformed the baseline, showing better performance when combining BERT with Word2Vec and RoBERTa with Fast-Text. These results clearly demonstrate that transformers are highly effective in capturing the small intricacies of medical sentiment. BERT with Word2Vec embeddings achieved the best accuracy of 95.71%, while RoBERTa with FastText achieved 94.47%. Future work might consider developing a dedicated dataset that would focus on medical sentiment analysis and extend its application to broader healthcare contexts.

## REFERENCES

[1] M. Mahyoob, J. Algaraady, M. Alrahiali, and A. Alblwi, "Sentiment Analysis of Public Tweets Towards the Emergence of SARS-CoV-2 Omicron Variant: A Social Media Analytics Framework," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8525–8531, Jun. 2022, https://doi.org/10.48084/etasr.4865.

[2] M. K. Myee, R. D. C. Rebekah, T. Deepa, G. D. Zion, and K. Lokesh, "Detection of Depression in Social Media Posts using Emotional Intensity Analysis," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16207–16211, Oct. 2024, https://doi.org/10.48084/etasr.7461.

[3] Y. Abdelwahab, M. Kholief, and A. A. H. Sedky, "Justifying Arabic Text Sentiment Analysis Using Explainable AI (XAI): LASIK Surgeries Case Study," *Information*, vol. 13, no. 11, Nov. 2022, Art. no. 536, https://doi.org/10.3390/info13110536.

[4] J. K. Lin, T. W. Chien, Y. T. Yeh, S. Y. C. Ho, and W. Chou, "Using sentiment analysis to identify similarities and differences in research topics and medical subject headings (MeSH terms) between Medicine (Baltimore) and the Journal of the Formosan Medical Association (JFMA) in 2020: A bibliometric study," *Medicine*, vol. 101, no. 11, Mar. 2022, Art. no. e29029, https://doi.org/10.1097/MD.0000000000029029.

[5] D. B. Fogel, "On the Growing Opportunity to Use Sentiment Analysis to Support Artificial Intelligence Applications in Healthcare," *Psychology and Behavioral Science International Journal*, vol. 10, no. 2, pp. 1–3, Jan. 2019, https://doi.org/10.19080/PBSIJ.2019.10.555798.

[6] T. Abbas *et al.*, "IoMT-Based Healthcare Systems: A Review," *Computer Systems Science and Engineering*, vol. 48, no. 4, 2024, https://doi.org/10.32604/csse.2024.049026.

[7] R. Alharbey, J. I. Kim, A. Daud, M. Song, A. A. Alshdadi, and M. K. Hayat, "Indexing important drugs from medical literature," *Scientometrics*, vol. 127, no. 5, pp. 2661–2681, May 2022, https://doi.org/10.1007/s11192-022-04340-7.

[8] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions," *IEEE Access*, vol. 11, pp. 90367–90382, 2023, https://doi.org/10.1109/ACCESS.2023.3307308.

[9] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 2016, https://doi.org/10.18653/v1/n16-1174.

[10] S. Xiong, H. Lv, W. Zhao, and D. Ji, "Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings," *Neurocomputing*, vol. 275, pp. 2459–2466, Jan. 2018, https://doi.org/10.1016/j.neucom.2017.11.023.

[11] J. Yu, J. Jiang, and R. Xia, "Global Inference for Aspect and Opinion Terms Co-Extraction Based on Multi-Task Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 168–177, Jan. 2019, https://doi.org/10.1109/taslp.2018.2875170.

[12] Z. Yuan, S. Wu, F. Wu, J. Liu, and Y. Huang, "Domain attention model for multi-domain sentiment classification," *Knowledge-Based Systems*, vol. 155, pp. 1–10, Sep. 2018, https://doi.org/10.1016/j.knosys.2018.05.004.

[13] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for Aspect-level Sentiment Classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, 2016, https://doi.org/10.18653/v1/d16-1058.

[14] I. Masood, Y. Wang, A. Daud, N. R. Aljohani, and H. Dawood, "Towards Smart Healthcare: Patient Data Privacy and Security in Sensor-Cloud Infrastructure," *Wireless Communications and Mobile Computing*, vol. 2018, no. 1, Jan. 2018, https://doi.org/10.1155/2018/2143897.

[15] P. Dumbach, L. Schwinn, T. Löhr, P. L. Do, and B. M. Eskofier, "Artificial intelligence trend analysis on healthcare podcasts using topic modeling and sentiment analysis: a data-driven approach," *Evolutionary Intelligence*, vol. 17, no. 4, pp. 2145–2166, Aug. 2024, https://doi.org/10.1007/s12065-023-00878-4.

[16] N. Ouerhani, A. Maalel, and H. Ben Ghézala, "SMAD: SMart assistant during and after a medical emergency case based on deep learning sentiment analysis: The pandemic COVID-19 case," *Cluster Computing*, vol. 25, no. 5, pp. 3671–3681, Oct. 2022, https://doi.org/10.1007/s10586-022-03601-7.

[17] H. Grissette and E. H. Nfaoui, "Deep associative learning approach for bio-medical sentiment analysis utilizing unsupervised representation from large-scale patients' narratives," *Personal and Ubiquitous Computing*, vol. 17, no. 6, pp. 2055–2069, Dec. 2023, https://doi.org/10.1007/s00779-021-01595-4.

[18] M. C. Martinis, C. Zucco, and M. Cannataro, "An Italian lexicon-based sentiment analysis approach for medical applications," in *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Northbrook, IL, USA, Aug. 2022, pp. 1–4, https://doi.org/10.1145/3535508.3545594.

[19] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artificial Intelligence in Medicine*, vol. 64, no. 1, pp. 17–27, May 2015, https://doi.org/10.1016/j.artmed.2015.03.006.

[20] D. C. Edara, L. P. Vanukuri, V. Sistla, and V. K. K. Kolli, "Sentiment analysis and text categorization of cancer medical records with LSTM," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 5, pp. 5309–5325, May 2023, https://doi.org/10.1007/s12652-019-01399-8.

[21] Y. Zhao, L. Zhang, C. Zeng, W. Lu, Y. Chen, and T. Fan, "Construction of an aspect-level sentiment analysis model for online medical reviews," *Information Processing & Management*, vol. 60, no. 6, Nov. 2023, Art. no. 103513, https://doi.org/10.1016/j.ipm.2023.103513.

[22] P. Kaur, A. K. Malhi, and H. S. Pannu, "Sentiment analysis of linguistic cues to assist medical image classification," *Multimedia Tools and Applications*, vol. 83, no. 10, pp. 30847–30866, Sep. 2023, https://doi.org/10.1007/s11042-023-16538-9.

[23] S. M. Jiménez-Zafra, M. T. Martín-Valdivia, M. D. Molina-González, and L. A. Ureña-López, "How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain," *Artificial Intelligence in Medicine*, vol. 93, pp. 50–57, Jan. 2019, https://doi.org/10.1016/j.artmed.2018.03.007.

[24] S. Sarkar, "Sentiment Analysis for Mental Health." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health.