NATURENGS
E-ISSN 2717-8013
https://dergipark.org.tr/en/pub/naturengs

A Deep Learning Framework for Detecting Depression Tendencies in Social Media Content

Sevda Güneyli [a], Serpil Aslan [a,*]

[a] Malatya Turgut Özal University, Department of Software Engineering, Malatya Türkiye – 44210

* Corresponding author

ARTICLE INFO

Received 16.05.2025
Accepted 23.06.2025

Doi: 10.46572/naturengs.1700871

ABSTRACT

This study aims to automatically detect depression-related content in social media posts. The Sentiment140 dataset was used, and the text data were preprocessed with natural language processing techniques and transformed into vector representations using GloVe embeddings. A balanced binary classification dataset was constructed by combining 15,000 positive tweets with 7,675 tweets containing depressive expressions. Classification was performed using machine learning and deep learning algorithms, and their performances were compared based on accuracy, precision, recall, and F1 score. The experimental results show that deep learning models, particularly the CNN architecture, outperformed traditional classification methods in detecting depression-prone content. CNN's success is attributed to its ability to capture local semantic patterns and process contextual features effectively. Overall, the study presents a practical and applicable approach for conducting mental health analysis based on social media data.

Keywords: Depression Detection, Social Media Analysis, GloVe, Deep Learning, Machine Learning, Text Classification, Sentiment140, Sentiment Analysis

1. Introduction

Depression is one of the most common mental health issues worldwide today. Depression is characterized by a variety of psychological and physical symptoms, including a constant state of sadness, loss of interest and pleasure, feelings of guilt and worthlessness, a lack of energy, sleep and appetite disorders, and attention and concentration problems [1]. These symptoms significantly limit the individual's daily life activities, functionality, and quality of life. If left untreated, depression can become chronic and lead to serious consequences such as suicide. Especially in low- and middle-income countries, due to structural difficulties in accessing mental health, a large proportion of depression cases remain undiagnosed, and individuals cannot receive professional help.

Studies show that approximately 75% of individuals in the early stages of depression do not receive any professional support, which causes the disease to progress [2,3]. This rate points to the limited access to mental health services at both individual and societal levels, and the stigmatization of mental problems is still a widespread problem. In this context, early diagnosis of depression, timely implementation of appropriate intervention strategies, and personalization of the treatment process are of great importance for the protection of mental health [4].

Traditional depression diagnosis methods are usually based on face-to-face interviews, scales, and clinical assessment tools conducted by psychiatrists or clinical psychologists. However, with the widespread use of digital technologies, how individuals express their emotional states has also changed significantly [5,6]. Social media platforms, in particular, have become a medium where individuals share their inner worlds, emotional states, and daily lives. These digital environments provide large volumes of naturally produced data, creating new opportunities for evaluating psychological states. Indeed, studies conducted in recent years have shown that content shared on social media platforms, especially Twitter, can provide meaningful clues about individuals' mental states [7,8,9]. Thus, social media data is a more dynamic, up-to-date, and natural data source than passive data collection methods.

Pioneering studies in this field have generally employed traditional machine learning methods to detect symptoms of depression in social media content. These approaches typically rely on classical algorithms such as Naive Bayes, Support Vector Machines (SVM), Decision Trees, and Logistic Regression [10]. However, the expression of depressive states in natural language tends to be subtle, context-dependent, and influenced by cultural and individual factors, which limits the effectiveness of these models in capturing underlying

* Corresponding author. e-mail address: serpil.aslan@ozal.edu.tr
ORCID : 0000-0001-8009-063X

semantic nuances and implicit cues. This limitation underscores the need for more sophisticated analytical methods in the detection of mental health conditions. In response to this need, the present study introduces a new dataset that combines a carefully selected subset of positive tweets from the widely used Sentiment140 corpus with a manually curated collection of tweets reflecting depressive expressions. This integrated dataset provides a realistic and balanced setting for training and evaluating binary classification models focused on distinguishing between general positive sentiment and depression-prone content.

Building on this foundation, the study compares the classification performance of traditional machine learning algorithms and deep learning-based natural language processing models. Through comprehensive evaluation, the study aims to identify the most effective approach for the automatic detection of depressive tendencies in social media posts. Ultimately, this research seeks to contribute to the development of accurate and scalable tools for early mental health screening and to support future digital mental health applications with robust empirical evidence.

2. Preliminaries

2.1. Machine learning models

Naive Bayes classifiers are probabilistic models based on Bayes' Theorem, commonly used in text classification, spam detection, and sentiment analysis. This method calculates the probability of belonging to a specific class, assuming that each feature to be classified is independent. After training with pre-labeled data, the model calculates the likelihood of a new sample belonging to specific classes and predicts the class with the highest probability. As the training data increases, so does the model's predictive power, allowing for more accurate classification of the test data. The Naive Bayes algorithm is a popular method for many applications due to its simplicity, low computational cost, and high scalability [11].

Logistic regression is a powerful statistical modeling technique used in binary or multiple classification problems. In particular, it estimates the probability of a dependent variable belonging to a particular class based on its relationship with one or more independent variables. Unlike linear regression, the model produces output through a sigmoid function that limits the predicted value between 0 and 1. In this respect, it is pretty effective in classification problems that involve limited and discrete results, such as "yes/no". Logistic regression is widely used for predictive analysis not only in the field of machine learning but also in many disciplines such as health, finance, and social sciences. This model is preferred in many practical applications due to its computational efficiency and high interpretability [12].

The Random Forest algorithm is an ensemble-based machine learning method based on multiple decision trees. In this method, each decision tree is trained independently on a different subset of the training data, and the majority of votes of these trees determines the final classification. Random Forest provides an advantage in balancing bias and variance issues. While there is a risk of overfitting when decision trees show high variance, the Random Forest structure minimizes this risk. In addition, when trees with low correlation are brought together, the model's generalization ability increases, and more accurate predictions are obtained. Providing high accuracy rates in classification and regression problems, the relative flexibility of parameter settings, and the ability to work with high-dimensional data make this method particularly useful in large data sets [13].

Decision tree algorithms are hierarchical models that perform classification and regression tasks by analyzing patterns in data within a logical structure. While each node represents a decision point based on a specific feature, leaf nodes contain the result classes. Decision trees are preferred, especially in applications where interpretability and visualization are important. This model, which is frequently used in operations research to develop decision support systems, clearly presents possible scenarios and probabilities [14]. It is also quite effective in the context of machine learning and can easily deal with data with a large number of features. However, since it can be vulnerable to over-learning alone, its performance is usually optimized with techniques such as pruning [15].

2.2. Deep learning models

Convolutional Neural Networks (CNNs) [16] are a robust artificial neural network architecture built on deep learning and explicitly designed for image and video data analysis. This architecture, developed by LeCun and his colleagues in the late 1990s, outperforms traditional fully connected networks in terms of learning efficiency and effectiveness by preserving the spatial structure of the data. CNN architectures are typically built with three basic components: convolution layers, pooling layers, and fully connected layers. Convolution layers filter local regions of an image to extract basic features like edges, textures, and shapes. The network's deeper layers transform these features into increasingly abstract and complex representations. Pooling layers, however, reduce the computational load and make the model robust against minor variations with dimensionality reduction operations. Fully connected structures in the last layers perform tasks such as classification or regression using the extracted features.

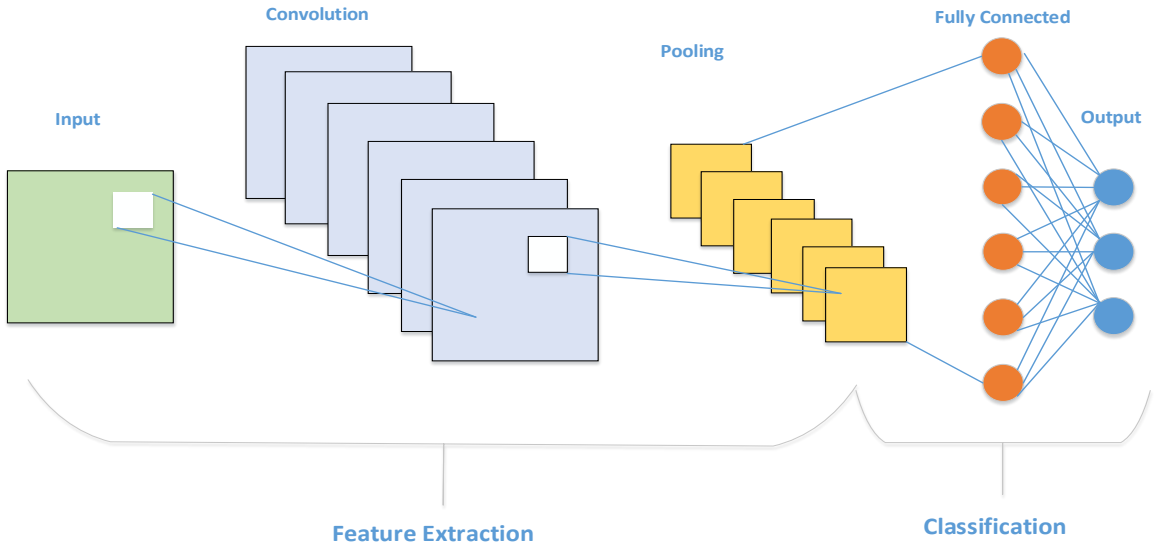

Figure 1. Basic Structure of the CNN Model

Long Short-Term Memory (LSTM) networks are a more advanced version of the Recurrent Neural Network (RNN) architecture, distinguished by their ability to model dependencies on sequential data. Traditional RNN structures have difficulty preserving past information, particularly when dealing with long sequences, and cannot learn long-term dependencies due to the "vanishing gradient" problem. LSTMs, which were created to overcome this structural limitation, have special cells and gate mechanisms that control the flow of information [17]. The LSTM cell has three basic components: the forget gate, the input gate, and the output gate. These mechanisms dynamically determine how much of the network will rely on previous information and what information will be updated. This structure allows LSTM models to retain important information over time while optimizing the learning process by removing irrelevant data. This capability provides significant advantages in sequential data analysis, including natural language processing, time series prediction, speech recognition, and machine translation. LSTMs are widely used in tasks that require complex context, such as language modeling, because they can process both short-term and long-term correlations at the same time. In this regard, LSTM is a viable option among deep learning approaches, prioritizing accuracy and contextual integrity.

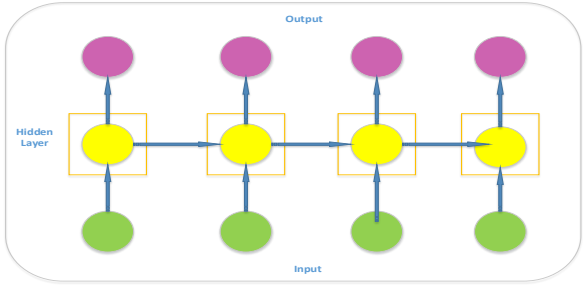

Figure 2. Basic Structure of the LSTM Model

Bidirectional Long Short-Term Memory (BiLSTM) networks are an advanced sequential data processing architecture consisting of two separate LSTM layers that work in opposite directions [18]. One part of this structure processes data from the past to the future, while the other processes the same inputs from the future to the past. Thanks to this bidirectional information flow, the model can simultaneously access contextual information from previous and subsequent time steps. BiLSTM architectures offer richer representation power by combining two-way hidden states. Thus, the model can make deeper inferences by considering past information and future correlations. This feature significantly increases performance, especially in applications with high context sensitivity, such as natural language processing. Bidirectional LSTM (BiLSTM) consists of two LSTM layers that process data forward and backward [18]. This bidirectional structure allows the model to learn past and future context simultaneously. Thus, a more accurate and robust representation is achieved, especially in tasks where context sensitivity is important.

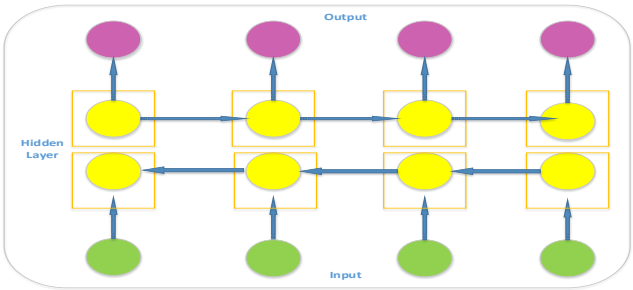

Figure 3. Basic Structure of the BiLSTM Model

2.3. Word embedding model

GloVe (Global Vectors for Word Representation) [19] is an unsupervised learning method based on a co-occurrence matrix that models semantic relationships between words with vector representations. The model aims to express the meaning numerically by using the co-occurrence statistics of words throughout the corpus.

Although the creation of the co-occurrence matrix is initially processor-intensive, the training process progresses more efficiently later on due to the sparsity of the matrix. Compared to local context-based models such as Word2Vec, GloVe offers more holistic representations by focusing on global relationships between words. In this respect, it produces effective results, especially in natural language processing applications where semantic closeness must be captured numerically.

3. The Proposed System

3.1. Dataset and preprocessing

In this study, a new dataset was created by combining a selected subset of the Sentiment140 dataset with a collection of tweets reflecting depressive expressions. The Sentiment140 dataset, which includes 1.6 million tweets labeled for sentiment polarity along with accompanying metadata such as text, user information, and timestamp, is frequently used in large-scale sentiment analysis due to its structured format and reliable annotations. From this corpus, 15,000 tweets labeled as positive sentiment were extracted. Additionally, 7,675 tweets containing depressive language were compiled to represent the negative class, resulting in a combined dataset of 22,675 tweets.

Within the study's scope, this dataset was restructured in the context of a binary classification problem to identify language patterns associated with depression. Since it does not contain a label directly defining depression, negatively labeled tweets in the data were evaluated as content with depressive tendencies; positive tweets were classified as examples with no symptoms of depression. This approach was adopted as an indirect classification strategy due to the lack of a publicly available dataset directly labeled for depression.

This dataset was designed for binary classification to distinguish between positive emotional expressions and language associated with depressive tendencies. As publicly available datasets labeled explicitly for depression are limited, an indirect classification strategy was adopted. In this context, positive tweets were assumed to reflect the absence of depressive symptoms, while the negative class was represented by tweets manually identified as having depressive content, despite the lack of clinical labeling.

Before model training, the dataset underwent a series of preprocessing steps to ensure compatibility with text mining and natural language processing methods. These steps included case normalization, removal of punctuation, elimination of stopwords, and deletion of non-semantic elements such as URLs and usernames. Following preprocessing, two separate word clouds were generated to visualize the most frequently occurring terms in each class. These visualizations, presented in Figure 4, offer a qualitative perspective on the lexical differences between the positive and depressive tweet groups and contribute to the interpretability of the model development process.

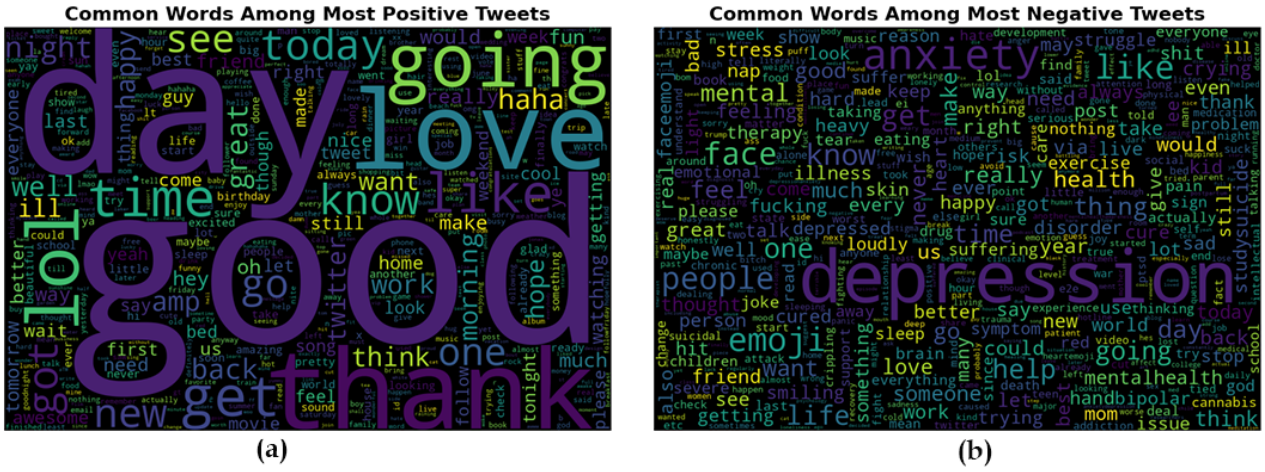

Figure 4. (a) Word cloud of tweets with depression tendencies, (b) Word cloud of tweets with positive content.

The word clouds in Figure 4 visually compare the prominent words in tweets labeled negative (a) and positive (b). In the word cloud of tweets with a tendency towards depression on the left side of the visual, it is seen that words indicating psychological distress and mental disorders such as “depression”, “anxiety”, “suicide”, “suffering”, “mental”, “alone” and “disorder” are prominent. These words reflect the intensity of the content in which individuals directly express their negative moods and psychological states. On the other hand, in the word cloud of positive tweets on the right side, words representing positive emotions and social interactions such as “good”, “thank”, “love”, “today”, “happy”, and “great” are prominent. This shows that tweets belonging to the positive class are primarily based on expressing gratitude, happiness, social bonds, and positive experiences related to daily life. This visual comparison reveals the differences in linguistic patterns between the two classes in the dataset on which the model was trained and supports the language-based underpinnings of the classification process.

3.2. The proposed model

This study proposes a method for automatically classifying tweets with a depression tendency. The process consists of three main stages: data

preprocessing, word embedding, and comparing classification models.

In the first stage, the texts in the Sentiment140 dataset were preprocessed with natural language processing techniques. In this context, uppercase letters were converted to lowercase letters; punctuation marks, links (URLs), usernames, and stopwords were cleaned from the texts. These processes aimed to make the data on which the model will be trained more consistent and meaningful.

In the second stage, the preprocessed data were converted into numerical vector representations with the GloVe (Global Vectors for Word Representation) method. The GloVe model learns the semantic relationships between words based on co-occurrence information and represents each word as a vector of a specific size. These representations were used as the input data for the classification models.

In the last stage, the obtained word vectors were applied to machine learning (LR, RF, DT, NB) and deep learning (BiLSTM, LSTM, CNN)- based classification algorithms. The performance of the models was analyzed using evaluation metrics such as accuracy, precision, recall, and F1 score. In this way, it was comparatively revealed which approach was more effective in classifying depression-prone content.

4. Experimental Results

In this section, the classification performance of the proposed approach is evaluated comparatively using various machine learning and deep learning algorithms. The dataset was split into 80% training and 20% test sets, and the evaluation metrics used include accuracy, precision, F1-score, and recall. The results are summarized in Table 1.

Among the deep learning-based models, CNN exhibited the highest overall performance with 99.03% accuracy, 99.07% precision, 99.38% F1-score, and 99.69% recall. The superior performance of CNN can be attributed to its capacity to learn position-invariant local features through convolutional filters. In text classification tasks, CNN excels at capturing discriminative n-gram patterns and relevant local dependencies within sentences, without requiring sequential input processing. Its ability to generate hierarchical feature representations makes it particularly powerful for extracting semantically rich information from textual data.

In this study, word embeddings were constructed using the GloVe (Global Vectors for Word Representation) method, which provides dense and semantically informative vector representations. The integration of GloVe embeddings with CNN significantly enhanced the model's ability to recognize contextually important word relationships, contributing to its high classification success. The combination of GloVe's semantic richness and CNN's feature extraction capabilities resulted in a robust architecture for handling complex textual patterns. BiLSTM followed closely in performance, achieving 99.12% accuracy, 99.28% precision, 98.70% F1-score, and 98.16% recall. BiLSTM is particularly effective in capturing long-range dependencies and bidirectional contextual information, allowing it to model the underlying semantic structure of text more comprehensively. LSTM also performed well, reaching 98.88% accuracy and 98.37% F1-score. Among the traditional machine learning algorithms, NB yielded the highest performance with 96.79% accuracy and 96.59% F1-score. RF achieved 90.76% accuracy and 95.12% F1-score. In contrast, SVM and LR displayed relatively lower classification performance, with accuracies of 88.72% and 85.12%, respectively. Overall, the experimental findings clearly demonstrate the superiority of deep learning models, particularly CNN and BiLSTM, in learning intricate linguistic structures and semantic relationships. The synergy between deep neural architectures and GloVe embeddings plays a crucial role in enhancing classification accuracy by enabling effective encoding of syntactic and semantic nuances within the data.

Table 1. The performance comparision of the classification models

Classification Model	**Performance Evaluation Metrics**			
	Accuracy ***(%)***	*Precision* ***(%)***	*F1-Score* ***(%)***	*Recall*
CNN	*99.03*	**99.07**	**99.38**	99.69
LSTM	98.88	98.25	98.37	98.48
BiLSTM	99.12	99.28	98.70	98.16
NB	96.79	88.41	96.59	92.32
RF	90.76	90.80	95.12	89.74
SVM	88.72	89.66	87.37	85.78
LR	85.12	81.84	82.00	88.31

5. Conclusion

This study presents a comprehensive approach for the automatic classification of tweets with depressive tendencies by integrating natural language processing techniques, semantic word embeddings, and state-of-the-art classification models. A new dataset was constructed by combining a subset of the Sentiment140 corpus with a manually curated collection of tweets reflecting depressive expressions, resulting in a balanced binary classification setting.

The preprocessing phase ensured that the textual data were clean and semantically consistent, while the use of the GloVe word embedding method enabled the transformation of linguistic input into meaningful vector representations. These representations captured both syntactic and semantic relationships, serving as a strong foundation for classification tasks.

Experimental findings indicate that deep learning models, particularly CNN and BiLSTM, significantly outperformed traditional machine learning approaches in identifying depressive content. CNN demonstrated the highest overall performance by effectively capturing local linguistic patterns and context-independent features, whereas BiLSTM exhibited strong competence in modeling long-term dependencies within text. Among traditional models, NB achieved the best results but remained behind the deep learning-based approaches in classification accuracy and consistency.

The results underline the potential of combining robust textual representations with advanced neural architectures to detect mental health indicators from social media content. Future studies may further improve classification reliability by incorporating larger and clinically annotated datasets, as well as leveraging multimodal data sources such as images, emojis, or interaction metrics to enrich emotional context.

Acknowledgements

This study was supported by the Scientific Research Projects Coordination Unit of Malatya Turgut Özal University under project number 25Y05.

References

[1] Tanzi, L., Vezzetti, E., Moreno, R., & Moos, S. (2020). X-ray bone fracture classification using deep learning: a baseline for designing a reliable approach. Applied Sciences, 10(4), 1507

[2] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., ... & Ng, A. Y. (2017). Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957

[3] Sahin, M. E. (2023). Image processing and machine learning-based bone fracture detection and classification using X-ray images. International Journal of Imaging Systems and Technology, 33(3), 853-865.

[4] Agarwal, D., Singh, V., Singh, A. K., & Madan, P. (2025). Dwarf Updated Pelican Optimization Algorithm for Depression and Suicide Detection from Social Media. Psychiatric Quarterly, 1-34.

[5] Kumar, M., Dredze, M., Coppersmith, G., & De Choudhury, M. (2015, August). Detecting changes in suicide content manifested in social media following celebrity suicides. In Proceedings of the 26th ACM conference on Hypertext & Social Media (pp. 85-94).

[6] Yan, Z., Peng, F., & Zhang, D. (2025). DECEN: A deep learning model enhanced by depressive emotions for depression detection from social media content. Decision Support Systems, 114421.

[7] Nguyen, S. D., Tran, T. S., Tran, V. P., Lee, H. J., Piran, M. J., & Le, V. P. (2023). Deep learning-based crack detection: A survey. International Journal of Pavement Research and Technology, 16(4), 943-967.

[8] Uysal, F., Hardalaç, F., Peker, O., Tolunay, T., & Tokgöz, N. (2021). Classification of shoulder x-ray images with deep learning ensemble models. Applied Sciences, 11(6), 2723.

[9] Tejaswini, V., Sathya Babu, K., & Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1-20.

[10] Amin, A. A., Iqbal, M. S., & Shahbaz, M. H. (2024). Development of intelligent fault-tolerant control systems with machine learning, deep learning, and transfer learning algorithms: a review. Expert Systems with Applications, 238, 121956.

[11] Reddy, K. B., Naidu, D. S. P., Deekshitha, N., & Srinivas, M. (2025, February). Exploring Hybrid Approaches for Sentiment Classification: A Comparative Study of LSTM, Naive Bayes, and Bayesian Network on IMDB Reviews. In 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE) (pp. 221-227). IEEE.

[12] Elkahwagy, D. M. A. S., Kiriacos, C. J., & Mansour, M. (2024). Logistic regression and other statistical tools in diagnostic biomarker studies. Clinical and Translational Oncology, 26(9), 2172-2180.

[13] Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.

[14] Wang, Z., & Gai, K. (2024). Decision tree-based federated learning: a survey. Blockchains, 2(1), 40-60.

[15] Deshpande, A., Dubey, A., Dhavale, A., Navatre, A., Gurav, U., & Chanchal, A. K. (2024, April). Implementation of an nlp-driven chatbot and ml algorithms for career counseling. In 2024 International Conference on Inventive Computation Technologies (ICICT) (pp. 853-859). IEEE.

[16] Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M., & Xie, Y. (2024). From cnn to transformer: A review of medical image segmentation models. Journal of Imaging Informatics in Medicine, 37(4), 1529-1547.

[17] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[18] Zhang, S., Zheng, D., Hu, X., & Yang, M. (2015, October). Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia conference on language, information and computation (pp. 73-78).

[19] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).