Journal of Computers and Intelligent Systems
Vol. 3, No. 4, 2025, pages 312 - 318
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

Language as a Lifeline: Leveraging NLP to Suicide Detection Through Context-Aware AI Models

Ghalib Nadeem[1], Dure Jabeen[2], Dilbar Hussain[2], Jamil Ahmed[2], Anees Ahmed[3], Abdul Khaliq[4], Alisha Shaikh[5]

[1]Department of Electrical and Computer Engineering, Iqra University, Karachi, Pakistan.
[2]Department of Computer Science, Iqra University, Karachi, Pakistan.
[3]Department of Computer Science, NEDUET, Thar, Pakistan.
[4]College of Computer Science and Information Systems, IoBM, Karachi, Pakistan.
[5]Department of Information Technology, Sydney, Australia.

ARTICLE INFO

Article History:

Received: October 01, 2025
Revised: November 05, 2025
Accepted: December 10, 2025
Available Online: December 20, 2025

Keywords:

Natural Language Processing
LSTM
Mental Health
BERT
TF-IDF

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

Each year, more than 700,000 people die of suicide, and this tragic phenomenon is the main cause of death in the world where mental health problems have become increasingly common (World Health Organization, 2018). One of the most significant opportunities for early detection of suicide is given by automated social media analysis, which recently has turned out to be a crucial outlet for people to express their feelings. This research analyzes the necessity of employing advanced natural language processing techniques to identify the intent of suicide in posts made on Reddit. It certainly has the potential to save lives. Three classifications are mentioned in the comparison as follows: a deep learning-based LSTM model, simple Logistic Regression using TF-IDF and a transformer-based BERT model. An accurately chosen dataset of Reddit articles which were evaluated for the risk of suicide was used to train and test models with standard metrics and 5-fold cross-validation. The results showed that BERT is by far the best of the alternatives. While LSTM gets 91% in terms of accuracy and score, Logistic Regression, on the other hand, gets only 87% for accuracy and 85% for F-score, thus performing poorly. The results of McNemar's test indicate the significance of BERT's superiority at $p < 0.05$ level. This research highlights the groundbreaking potential of context-aware language models in diagnosing mental health conditions. The adoption of these approaches on digital platforms will benefit all parties involved, such as physicians, researchers, and technology firms, to facilitate real-time, scalable, and ethical surveillance of suicidal conduct. In an era where every signal is essential our work represents a crucial advancement in AI-assisted mental health solutions.

Corresponding Author's Email:

Citation:

1. Introduction

The worldwide mental health problem is growing significantly, with suicide now ranked as the fourth greatest cause of death among those aged 15 to 29 worldwide. The World Health Organization reports that more than 700,000 individuals died because of suicide annually, equating to one death every 40 seconds [1]. There are over 20 suicide attempts and numerous individuals contend with suicidal thoughts highlighting it as most important problem. The cost in addition to personal suffering is immense. Suicide and self-harm around $93.5 billion per year in medical and productivity losses in the United States [2]. While progress in mental health treatment, traditional strategies to recognizing suicidal ideation, such as patient self-reporting and clinician intervention continue to be reactive constrained in scalability and frequently delayed. These methods often overlook the essential period for early intervention.

In addition, the rapid development of social media sites such as Facebook, Twitter and Reddit has changed the nature of human interaction because users more likely to broadcast their innermost feelings in real time. Notably, Reddit serves as a major forum where people may openly discuss mental health issues, including thoughts of suicide. Reddit is a vast forum that attracts an average of 52 million users daily [3]. It constitutes an enormous collection of online emotional and behavioral signs of individuals on a large scale. The digital footprint formed here presents an exceptional chance to identify early signs of mortality through an automated, real-time monitoring system, in contrast to conventional in-person procedures. One subfield of artificial intelligence where machines analyze and interpret human languages is called natural language processing (NLP). NLP techniques have already proven successful in detecting mental health indicators such as depression, anxiety, and substance use problems by finding patterns that indicate negative sentiment, hopelessness, social withdrawal, or ideation of self-harm. NLP models could hence be robust tools for the assessment of suicide risks at an early stage. For instance, researchers find that individuals at high risk for suicide use first-person singular pronouns ("I," "me") significantly more often than average and show cognitive distortions in their languages.

In the earlier times, Bag of words and Term Frequency- Inverse Document Frequency (TF-IDF) techniques were learned with the help of classifiers like a Support Vector Machine (SVM) or Logistic Regression. These are useful, but they turn out to be context-insensitive. With the introduction of new deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM), modeling languages improved as these were focused on the order or sequence of words in the language. But long-range dependencies are still a challenge to the RNNs and LSTMs despite their many advantages. Transformer-based Bidirectional Encoder Representation from Transformers (BERT) models, like Bidirectional Encoder Representations from Transformers, completely changed the state-of-the-art. BERT is a bi-directional language model that processes inputs in both the forward and backward directions. BERT pretrained on extensive datasets and subsequently fine-tuned for certain tasks, has attained state-of-the-art performance in NLP applications, such as mental health detection. Its language understanding is particularly advantageous for detecting suicidal ideation that conventional models might over look.

This paper fills a significant research gap by evaluating three NLP models: Logistic Regression with TF-IDF, LSTM networks and BERT utilizing a Reddit dataset for the detection of suicidal thoughts. To assess the models, several rigorous evaluation metrics, including recall, precision, McNemar's test and F1-Score accuracy could be used. The digital behavior increasingly reflects emotional well-being, utilizing AI-driven solutions for suicide detection is not merely a technological achievement; it is a moral obligation. With accurate, scalable, and ethically deployed NLP systems, it can shift from reactive to proactive mental health care and, ultimately, save lives. The remaining paper is structured as follows: Section II is about the related work, Section III represents the methodology, and Section IV is about the results with a discussion. The V section is the conclusion and future work.

2. Related Work

Detection of suicidal ideation through digital text analysis is now an important subset of Natural Language Processing (NLP) for detecting mental health conditions. With the growth of social media platforms, Reddit and Twitter have become important data resources for early detection models since user-generated content tends to be freely expressed. According to Arowosegbe and Oyelade (2023), NLP has shown effectiveness in the prediction modeling of depression and Post-Traumatic Stress Disorder (PTSD) through the analysis of structured and unstructured data [4]. Communication of emotional cues with text could support intervention strategies for early mental health conditions. In the same way, other studies apply sentiment analysis to study suicide-related discourse using a variety of analytic approaches, including BERTopic and network-based analysis [5].

The early attempts in NLP extended towards lexicon-based or statistical modelling, employing TF-IDF together with Logistic Regression or SVM [6]. These models were interpretable rather than contextual and did not capture most nuances of emotional language [7]. The integration of such deep learning models as LSTM networks indicated a form of sequential dependencies capture, but even so, LSTM models exhibited major weaknesses regarding long-distance relationships and subtle context. Significantly, transformer-based models, particularly featuring BERT as a breakthrough that also allowed understanding a given context in both directions of bidirectional reading. Aside from that, BERT showed great advances for most emotion classification, sentiment analysis, and other tasks like depression detection [8]. Research employing BERT on Reddit data has regularly yielded robust outcomes. For example, [9] employed a temporal attention-enhanced BERT model, attaining an F1-score of 0.91. Likewise, [10] exhibited RoBERTa's cross-platform proficiency with a performance of roughly 0.90.

Ensemble methodologies have also garnered prominence. The CLPsych shared job, as reported by [11], integrated transformers with manually built features, resulting in exceptional F1-scores of 0.92. Nevertheless, these models frequently sacrifice simplicity and scalability for minimal performance improvements. Despite robust individual model outcomes, numerous existing research either fails to conduct comparison evaluations across modeling methodologies or neglect statistical validation. This study systematically compares three distinct models: Logistic Regression, LSTM, and BERT, using a dataset of 10,000 Reddit posts. The performance of BERT has been put to the test with McNemar's test as shown in Table 1, which contrasts strategic and evaluative outcomes of present research. This study looks into the development of suicide risk detection shifting from rule-based models to transformer architectures. Our research builds upon this groundwork by providing a BERT implementation that is backed by rigorous testing and is of high performance, thus indicating its appropriateness for large-scale and real-world applications.

Table 1. NLP Models for the Detection of Suicidal Ideation: A Survey (2020–2025)

Ref:	NLP Models comparison			
	Comparison	**Dataset**	**Models**	**Results**
7	Our LSTM achieves an F1-score of 0.90	Reddit (Suicide Watch, depression)	Logistic Regression, LSTM	LSTM F1 ≈ 0.86
8	Our BERT outperforms the F1 score of 0.93	Twitter	BERT, Logistic Regression	BERT F1 ≈ 0.89
9	Slight improvement In our study	Reddit + Temporal Patterns	BiLSTM, BERT w/attention	BERT F1 ≈ 0.91
11	Ours matches top-tier performance	Reddit (multi-subreddit)	LSTM, Transformer, Ensemble	Best F1 ≈ 0.92
10	Comparable to our BERT F1 score of 0.93	Reddit + Twitter	RoBERTa, SVM	RoBERTa F1 ≈ 0.90
Current Study	Statistically validated, highest F1 score	Reddit (10,000 labeled posts)	Logistic Regression, LSTM, BERT	BERT F1: 0.93, Acc: 94%

BERT, being the one to process data, used a collection of Reddit posts that summed up to 10,000 in numbers. The performance of BERT was established through the application of the McNemar's test and Table 1 by providing an exhaustive comparison of strategic and evaluative outcomes from present-day studies. Following the aforementioned, the path of the suicide risk detection technology has been shown to move from traditional methods to transformer models. By adding a high-performing BERT implementation that has been thoroughly tested, our work strengthens this foundation and confirms its viability for large-scale, real-world applications.

3. Methodology

In this part, we detail the methodology that our study used to identify postings on Reddit that had suicide thoughts by means of natural language processing. Information about the dataset, its preparation, feature extraction, model designs, training methods, and assessment criteria. Logistic Regression, LSTM, and BERT are three separate modeling paradigms that were purposefully included in the technique to guarantee a thorough, equitable, and repeatable comparison. Suicidal ideation detection methods flow steps are illustrated in Figure 1.

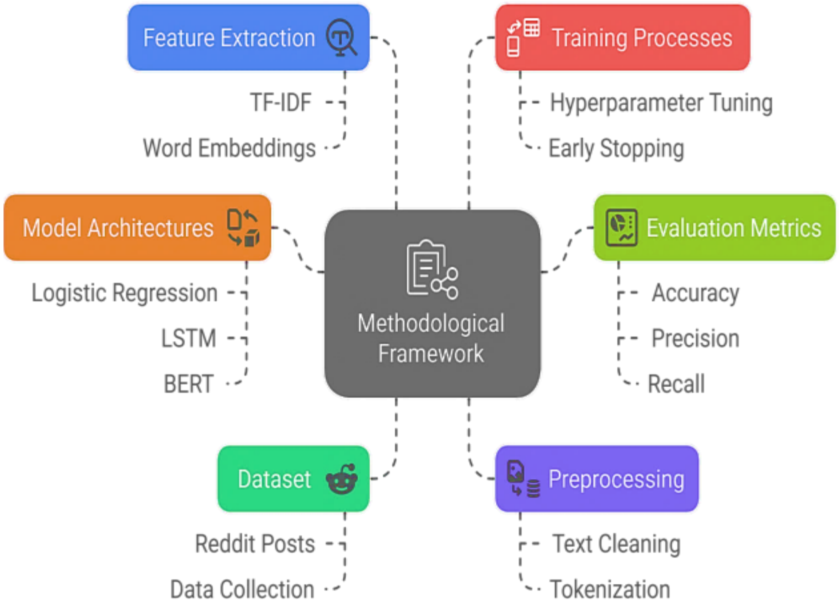

Figure 1. Methodology Framework for the Suicidal Ideation.

A. Dataset Description

The database that is being utilized in this work has been harvested from the Mendeley repository [12], "Suicide Risk Detection in Reddit Posts". It gathers thousands of Reddit posts marked either "suicidal" or "non-suicidal." These posts were gathered from many Subreddits related to mental health and represent a diverse range of language by those suffering mental distress. The data is balanced so as not to bias the models, but rather condition them both in training and evaluation. Every post is assigned a binary label, with "1" indicating suicidal and "0" indicating non-suicidal. This binary classification arrangement not only reduces the difficulty of the exercising task but also focuses equally on the main point in suicide risk detection.

B. Text Preprocessing

Pre-processing is an essential step in the natural language processing pipeline when processing unstructured user-generated content from sites like Reddit. Text pre-processing involves a series of procedures to deliver a clean and uniform input for all models. To bring the content to a common standard, it is first converted to lowercase and thereafter the redundancies are removed such as URLs, HTML tags, digits, and other special characters. Finally, the stop words like "the," "is," and "and" are removed as they do not carry any semantic meaning. In order to reduce the size of the vocabulary, we first used tokenization to break the text down into individual words and then lemmatization to bring the words back to their base forms. The occurrence of emotionally charged words like "want," "know," "feel," "kill," and "life" pointed out the intensity and frequency of mental health-related discussions, as illustrated in the Figure 2. This type of language indicates the importance of context-aware algorithms in detecting suicide risk through text data. Libraries like NLTK and spaCy were used for the execution of these phases, producing text that was uniform and ready for effective model training.

C. Modeling Approaches

1) Regression with TF-IDF:

The baseline model in our study is Logistic Regression using TF-IDF vectorization. TF-IDF transforms the text into a numerical representation by capturing how important a word is to a document in a corpus [13, 14]. Vocabulary is limited to a size of 5000 words, and unigrams and bigrams are applied to capture short phrases. Logistic Regression was selected for its simplicity and interpretability. Hyperparameters such as regularization strength were tuned using grid search on a validation set.

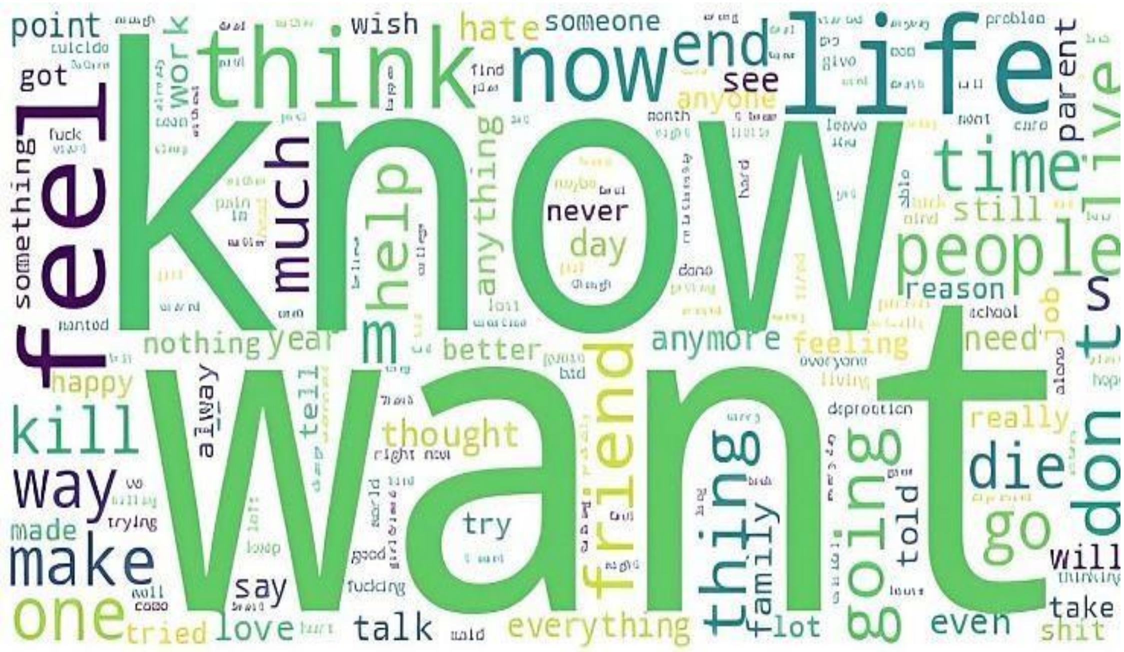

Figure 2. Word cloud visualization of frequently occurring terms in Reddit posts associated with suicidal ideation.

2) LSTM Model:

The LSTM network is a type of RNN capable of learning long-term dependencies in sequential data [15]. The LSTM model architecture consisted of an Embedding layer with a vocabulary size of 10,000 and an embedding dimension of 128, followed by an LSTM layer with 64 units and dropout regularization to prevent overfitting. A Dense output layer with a sigmoid activation function was used for binary classification. Input sequences were padded to a maximum length of 200 tokens to ensure uniform input shape. The model was trained using binary cross-entropy loss and the Adam optimizer with a learning rate of 0.001. Training was performed over 5 epochs, with early stopping applied based on validation loss to prevent overfitting.

3) BERT Model:

The BERT model represents the most advanced architecture in our comparison [16]. Working on our data, which was fine-tuned using the BERT-base-uncased model from Hugging Face, had been fine. The pre-processing for BERT-based modeling, tokenization, was carried out using the WordPiece tokenizer, with the addition of special tokens like [CLS] and [SEP]. Input sequences were then padded, and attention masks were generated to distinguish real tokens from padding. The BERT architecture used in this study comprised 12 transformer layers, 12 attention heads, and approximately 110 million parameters. Fine-tuning was carried out using a batch size of 16, a learning rate set to 2e-5, and three epochs. The output of the [CLS] token was passed through a linear layer to produce binary classification labels by applying a sigmoid activation function.

D. Training and Evaluation

To maintain the same equality and reliability across all models, training and evaluation were performed on the same 5-fold cross-validation scheme. The selected metrics for model performance include accuracy (overall proportion of correct predictions), precision (the proportion of positives predicted that are true positives), recall (the proportion of actual positives correctly detected), and F1 score (the harmonic mean of precision and recall). McNemar's test was also performed for statistical comparisons of model predictions and their significance. Hyperparameter tuning was performed using a validation split from the training data with early stopping and dropout applied to avoid overfitting. All the experiments were run in a GPU environment with TensorFlow and PyTorch libraries.

4. Results and Discussion

The outcome of our studies gives a detailed specification of detecting suicidal ideations from Reddit posts across each model. All three models were evaluated and compared based on accuracy, precision, recall, F1, and statistical significance via McNemar's test. The consistency of the experiment sets such that any differences noted must be due to model architecture and not external biases or inconsistency in data treatment. All models were trained and tested under 5-fold cross-validation to ensure generalization and reduce the effect of data-specific variance. The averaged results through the folds are reported in Table 2.

TABLE 2. RESULT COMPARISON

Model	Model Performance Metrics (Averaged over 5 folds)			
	Accuracy	Precision	Recall	F1-Score
Logistic Regression	87	86	84	85
LSTM	91	90	89	90
BERT	94	93	92	93

Though it's simple and easy to use and interpret, the Logistic Regression showed the lowest performance: it gave 87% accuracy and an F1-score of 85%. Because this method involves learning contextual meaning, it fails to recognize subtle signs of suicidal ideation. However, the LSTM model yielded a much better output at 91% accuracy and a 90% F1-score. By this means, the model can identify and take into account the temporal and emotional patterns in a specific text for the occurrence of distress cues. However, it sometimes stutters on longer or syntactically complex posts where the context from distant words has faded. Compared to others, BERT showed a significant performance, achieving 94%, with an F1-score of 93. Its bidirectional understanding has made it master both the very subtle and ambiguous languages very efficiently, as traditional and recurrent models usually lack competency in such areas. This underscores the impressive power of transformer-based models, particularly BERT, in the area of mental health detection. BERT's power comes from its vast understanding of the language and goes deep into the context which is the very thing needed to detect ts subtle, indirect, or even humorous ways of expressing suicidal intent that are often found on social media platforms like Reddit. However, a nice, high accuracy still is not without its ethical dilemmas when it comes to the real-world application of these models where privacy, transparency, and user consent come into play. Without checks and balances, including human oversight and clear data regulations, such systems are vulnerable to misuse. The issue of false positives and false negatives is also a major concern. An incorrect categorization of a post could cause unnecessary concern or, more seriously, impede a request for help. Human evaluation and clinical validation are critical; soaring to 94 percent accuracy, AI should augment expert judgment rather than replace it.

Other issues that need to be addressed are accessibility and equality. Due to the cultural differences, the models created based on the English-language Reddit data might not have the desired generalizability. The performance comparison of the languages will require long and tedious training on multiple languages and platforms. These obstacles also provide a very good chance to move further in the field. The use of real-time social media monitoring with natural language processing for the delivery of information that is traditionally provided by mental health providers in an inadequate, late, and non-scalable way. When used properly, these technologies can act as early warning systems, informing users of potentially dangerous information and allowing them to take life-saving actions.

5. Conclusion and Future Work

This research is going to showcase the superior performance of an NLP model such as BERT, together with its technical capabilities proof in recognizing suicidal intent in Reddit messages. BERT has a 94% accuracy rate and a 93% F1-score which is higher than both Logistic Regression (the best classical machine learning model) and LSTM (the best deep learning model) indicating the prominence of contextual language understanding in mental health applications. As a result, BERT might be utilized in the area of practical suicide detection. The emergence of new modes of emotional expression, particularly through social media, calls for the implementation of ethical monitoring mechanisms. This study has however raised concerns over data privacy, bias, and the necessity of human supervision, but it also marks a very significant milestone in the use of AI in the mental health sector thus, giving a stronger push to AI-based solutions for a critical public health issue. NLP models, such as BERT, hold a lot of promise for the early detection of suicide risk and the provision of timely intervention support.

In the future, the main points of research will revolve around the improvement of datasets for multimodal representation, multilingual content, and the application of natural language processing (NLP) models in clinical settings and mental health support systems. Explainable AI facilitates the confidence of both users and doctors in AI-influenced decision-making. The principal objective is to strengthen human interaction instead of replacing it, making sure that no patient is ever denied help by the doctor and that the doctor is always willing to render assistance. The results of the research can provide directions for creating mental health interventions that are flexible, ethically acceptable, and scalable in terms of their impact. Future research will concentrate on multi-label classification for recognizing more mental health disorders, real-time system deployment and monitoring for timely intervention, cross-platform analysis of social media domains like Twitter and Facebook for extensive insights, and the establishment of moral principles for guaranteeing responsible and reliable AI-assisted suicide prevention. Advancement in these areas will also bring us closer to the future of intelligent and compassionate systems that can effectively render the global mental health agenda.

6. References

1. World Health Organization (WHO), https://www.who.int/publications/i/item/9789241564779
2. Peterson, C., Haileyesus, T., & Stone, D. M. (2024). Economic cost of US suicide and nonfatal self-harm. American journal of preventive medicine, 67(1), 129-133.
3. Corbet, S., Hou, G., Hu, Y., & Oxley, L. (2021). We reddit in a forum: The influence of messaging boards on firm stability. Available at SSRN 3776445.
4. Arowosegbe, A., & Oyelade, T. (2023). Application of natural language processing (NLP) in detecting and preventing suicide ideation: a systematic review. International Journal of Environmental Research and Public Health, 20(2), 1514.
5. Kim, K., Kogler, D. F., & Maliphol, S. (2024). Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTopic. Humanities and Social Sciences Communications, 11(1), 1-15.
6. Srivastava, R., Bharti, P. K., & Verma, P. (2022). Comparative analysis of Lexicon and machine learning approach for sentiment analysis. International Journal of Advanced Computer Science and Applications, 13(3), 71-77.
7. Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., & Resnik, P. (2021). Expert, crowd, and machine: A meta-analysis of suicidal ideation detection on social media. Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, 123–132. https://doi.org/10.18653/v1/2021.clpsych-1.13
8. Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 54(8), 5789- 5829.
9. Fraga, B. S., da Silva, A. P. C., & Murai, F. (2018, December). Online social networks in health care: a study of mental disorders on Reddit. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 568-573). IEEE.
10. Buddhitha, P., & Inkpen, D. (2023). Multi-task learning to detect suicide ideation and mental disorders among social media users. Frontiers in research metrics and analytics, 8, 1152535.
11. Tsakalidis, A., Chim, J., Bilal, I. M., Zirikly, A., Atzil-Slonim, D., Nanni, F., ... & Liakata, M. (2022, July). Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology (pp. 184-198).
12. Mafi, Md Mafiul Hasan Matin; Alam, Md. Sabbir (2023), “Suicidal Ideation Detection Reddit Dataset”, Mendeley Data, V2, doi: 10.17632/z8s6w86tr3.2
13. Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, Tf- Idf, Word2vec and Doc2vec. SLU Journal of Science and Technology, 4(1), 27-33.
14. Ahmed, H., Haque, M. F. U., Khan, H. R., Nadeem, G., Arshad, K., Assaleh, K., & Santos, P. C. (2024). Selecting the best compiler optimization by adopting natural language processing. IEEE Access.
15. Vennerød, C. B., Kjærran, A., & Bugge, E. S. (2021). Long short-term memory RNN. arXiv preprint arXiv:2105.06756.
16. Alaparthi, S., & Mishra, M. (2020). Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. arXiv preprint arXiv:2007.01127.