International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 15 Issue 3, March-2026

A Real-Time Explainable AI System for Suicidal Expression Classification on Social Networking Platform

Sakhu Narhe[1], Dr. Sujeet More[2], Dr. Geetika Narange[3], Dr. Swati Mohite[4]

[1,2,3] Computer Engineering Trinity College of Engineering & Research, Pune, India

ABSTRACT

Mental health is being discussed more openly than ever, especially across social media. Because of this, there is a growing need for tools that can help identify when someone may be experiencing suicidal thoughts. To address this, we developed a system that analyzes social media text to detect signs of mental distress. The system is built using an improved version of DistilBERT and implemented with PyTorch. It collects real-time data from social media through Apify, ensuring timely and relevant input. Before analysis, the data is carefully cleaned and processed to maintain quality and accuracy. To promote transparency and trust, the system includes explainability features using SHAP, which help clarify how predictions are made. The final model is deployed through a Streamlit interface, allowing for real- time, accessible, and responsible mental health monitoring.

Keywords—Suicidal ideation detection, social media analysis, natural language processing (NLP), transformer models, DistilBERT, deep learning, text classification, explainable AI, SHAP, mental health monitoring

INTRODUCTION

Social media has transformed the way people communicate, share experiences, and express their emotions. Platforms such as Instagram, Twitter, Reddit, and Facebook have become spaces where users openly talk about personal struggles, mental health concerns, and emotional pain. While this openness has helped increase awareness and foster peer support, it has also led to a rise in sensitive and vulnerable content, including posts related to self-harm and suicidal thoughts. Suicide continues to be one of the leading causes of death among young people worldwide, highlighting the urgent need for early identification and timely support. However, the massive volume and rapid flow of content shared every day make manual monitoring impractical and unsustainable.

Recent advances in natural language processing (NLP) and deep learning have shown strong potential for automating the analysis of large-scale textual data. Transformer-based architectures, particularly BERT and its variants, have achieved state-of- the-art performance across a wide range of NLP tasks, including sentiment analysis, topic classification, and emotion detection. These models learn rich contextual representations of text, allowing them to capture subtle semantic patterns linked to emotional and psychological states. However, despite their high accuracy, many of these models remain difficult to interpret, limiting insight into how predictions are formed.Most transformer-based systems function as black-box models, offering little insight into the reasoning behind their predictions. This lack of transparency presents serious challenges in sensitive application areas such as mental health, where interpretability, trust, and accountability are essential.

In the context of suicidal ideation detection, misclassification can have serious consequences. False negatives may lead to missed opportunities for timely intervention, while false positives can cause unnecessary distress or stigmatization. As a result, there is a critical need for systems that not only deliver high predictive accuracy but also provide transparent and interpretable explanations for their decisions. Explainable Artificial Intelligence (XAI) methods, such as SHAP (SHapley Additive exPlanations), address this need by quantifying the contribution of individual features or tokens to a model's predictions.

This paper presents an end-to-end framework for the accurate and transparent classification of suicidal expressions on social media platforms. The proposed system is designed to operate on real-time data and supports the complete processing pipeline, including data extraction, preprocessing, analysis, and classification of user-generated content. At its core, the framework employs a fine-tuned DistilBERT model—an efficient transformer architecture that maintains strong performance while reducing computational complexity. The model is implemented using PyTorch and integrated

into a modular pipeline that encompasses data collection, text preprocessing, tokenization, inference, and result visualization.

To support real-world deployment, the system integrates automated data collection through the Apify API, enabling large- scale extraction of Instagram comments based on user-defined inputs. This approach allows the framework to process live data rather than relying on static datasets, making it well suited for real-time monitoring scenarios. The preprocessing pipeline normalizes, cleans, and validates raw text before it is passed to the model, enhancing robustness and generalization.

Beyond classification performance, the framework places strong emphasis on transparency and user trust. An explainability module based on SHAP is incorporated to identify the words and phrases. This functionality helps mental health professionals, content moderators, and researchers better understand the system's decision-making process, supporting ethical and responsible use. Additionally, the system is deployed through a Streamlit-based web interface, enabling real-time interaction and immediate feedback for users

LITERATURE REVIEW

The main contributions of this research are as follows:
(i) the design of a scalable and modular architecture for detecting suicidal expressions in social media content;
(ii) the integration of a fine-tuned DistilBERT model to achieve accurate classification;
(iii) the incorporation of explainable AI techniques to improve transparency and interpretability; and
(iv) the development of a real-time deployment framework using Streamlit. By combining strong performance, explainability, and practical usability, the proposed system helps bridge the gap between research prototypes and real-world mental health monitoring solutions.

The creating new opportunities for identifying mental health conditions such as suicidal ideation. As a result, researchers have increasingly focused on analyzing linguistic and emotional patterns in online content to support early detection and prevention of self-harm.

1] Early research in this area primarily relied on traditional machine learning approaches, using handcrafted textual features such as TF-IDF, sentiment polarity, and curated psychological lexicons to classify suicidal and non-suicidal expressions. While these methods produced encouraging results, their performance was limited by shallow contextual understanding and dependence on manually engineered features, which often failed to capture subtle emotional cues and complex semantic relationships in natural language.

2] With advances in deep learning, neural network-based models such as CNNs and RNNs became prominent in suicidal ideation detection research.

3] Architectures like LSTM and GRU were particularly effective at modeling sequential dependencies in social media text, enabling better recognition of emotional progression and distress signals. Several studies reported notable improvements over traditional classifiers, especially when large annotated datasets were available.

4] However, these models struggled with long-range contextual dependencies and were often computationally demanding, limiting their suitability for real-time analysis of large-scale social media data.

5] The introduction of transformer-based architectures marked a major breakthrough in natural language processing for mental health analysis. Models such as BERT, RoBERTa, and DistilBERT use self-attention mechanisms to capture bidirectional contextual information, allowing for more accurate interpretation of nuanced expressions of despair, hopelessness, and suicidal intent.

6] Recent studies consistently show that transformer-based models outperform both classical machine learning and recurrent neural network approaches in suicidal ideation classification tasks.

7] DistilBERT has gained particular attention due to its reduced size and faster inference speed, while maintaining competitive performance, making it well suited for real-time monitoring applications.

8] Despite their strong predictive performance, transformer-based models are often criticized for their lack of transparency. In sensitive domains such as mental health, the black-box nature of these models raises ethical concerns related to accountability, trust, and decision justification.

9] To address these challenges, researchers have increasingly incorporated explainable AI techniques into suicidal ideation detection frameworks. Methods such as SHAP and LIME have been used to generate post-hoc explanations by identifying influential words or phrases that contribute to model predictions.

10] Studies combining SHAP with transformer models suggest that explainability improves user trust and supports a deeper understanding of model behavior, which is essential for responsible and ethical deployment in mental health applications.

Another limitation observed in existing research is the heavy focus on offline evaluation rather than real-world deployment.

Most studies rely on static datasets collected from platforms such as Twitter and Reddit and do not address the challenges associated with continuous data streaming, live preprocessing, or system scalability. Additionally, there has been limited emphasis on developing end-to-end frameworks that integrate data collection, classification, explanation, and visualization within a single deployable system. This gap highlights the need for research that not only improves predictive accuracy but also prioritizes interpretability, computational efficiency, and real-time usability.

Overall, the literature shows a clear progression from traditional machine learning to deep learning and transformer-based approaches for suicidal ideation detection. While recent models have achieved strong performance, challenges related to transparency, ethical deployment, and real-time applicability remain insufficiently addressed. These limitations motivate the development of integrated, explainable, and efficient systems capable of accurately classifying suicidal expressions on social networking platforms while ensuring interpretability and practical relevance.

PROBLEM STATEMENT

The increasing expression of suicidal thoughts on social networking platforms presents a significant challenge for timely mental health intervention. Although automated text classification methods have been developed to detect suicidal content, many existing approaches rely on handcrafted features or complex deep learning models that lack contextual understanding, interpretability, and real-time applicability. Transformer-based models improve detection accuracy, but they often function as black-box systems, limiting transparency in sensitive mental health settings. Furthermore, most studies focus on offline evaluation and do not address deployment challenges such as continuous data processing and explainable decision-making. Therefore, there is a critical need for an accurate, efficient, and interpretable framework that can classify suicidal expressions from social media text in real time while providing transparent and trustworthy predictions suitable for practical deployment.

OBJECTIVES

- **Design an automated framework** for accurately classifying suicidal and non-suicidal expressions from social media text.
- **Leverage transformer-based language models** to capture contextual and semantic features associated with suicidal ideation.
- **Integrate explainable AI techniques** to provide transparent and interpretable model predictions.
- **Ensure computational efficiency and scalability** for real-time or near real-time deployment.
- **Evaluate the proposed framework** using standard performance metrics and compare it with existing baseline approaches.

METHODOLOGY

Section 2 reviews related work in suicidal ideation detection and explainable NLP systems. Section 3 presents the proposed system architecture, while Section 4 details the methodology, including data collection, preprocessing, and model training. Section 5 discusses the evaluation metrics and experimental results. Section 6 introduces the explainability framework, and Section 7 concludes the paper and outlines directions for future research.

Fig. Instagram Comment Analysis Interface

This figure illustrates the user interface for analyzing Instagram post comments. The user inputs an Instagram post URL and selects the number of comments to retrieve. The system confirms API configuration and enables users to start comment extraction and analysis with a single action.

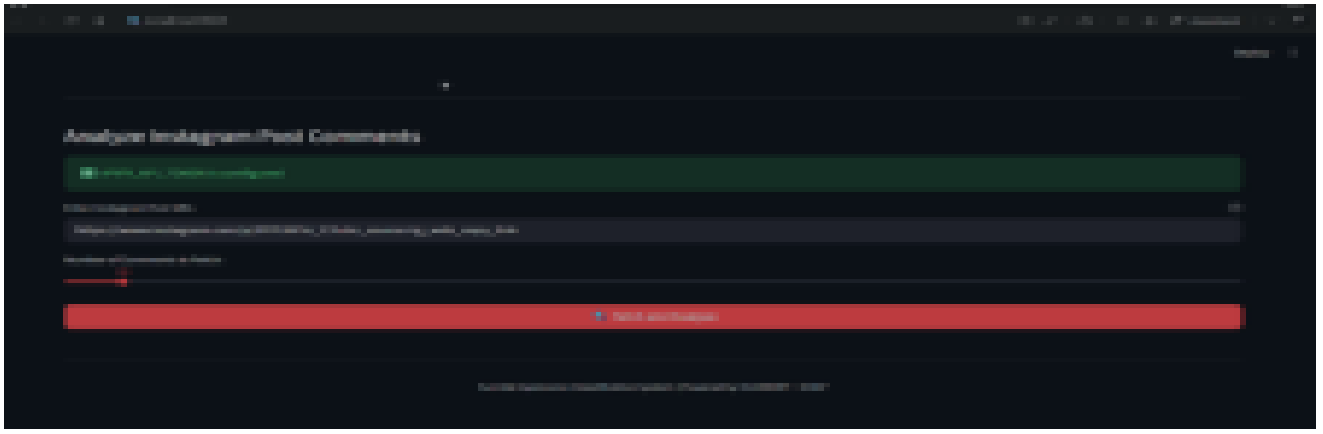

Fig. Suicide Expression Classification Dashboard

This figure shows the main dashboard of the suicide expression classification system. It confirms successful model

loading and allows users to choose between direct text input or Instagram link–based analysis, demonstrating the system's flexibility in handling multiple input sources.

Fig. Comment Fetching and Analysis Status with Results Summary

This figure illustrates the system's execution stage, showing the successful retrieval of Instagram comments along with validation and debugging information. It also provides a summarized view of the classification results, including the total number of analyzed comments and the counts of predicted suicidal and non-suicidal expressions, demonstrating the system's end-to-end processing capability.

Fig. Classification Results Display

This figure shows the classification results in a tabular format, summarizing the total number of analyzed comments along with the counts of suicidal and non-suicidal expressions. Each comment is displayed with its predicted class and confidence score, supporting transparent and interpretable analysis outcomes.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[3] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP*, 2020, pp. 38–45.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[7] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[8] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12,

[9] pp. 2825–2830, 2011.

[10] A. Paszke et al., "PyTorch: An imperative style, high performance deep learning library," in *Advances in*

Neural Information Processing Systems, 2019.
[11] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.
[12] J. Lin et al., "Detecting suicidal ideation in social media: A deep learning approach," *IEEE Transactions on Affective Computing*, 2020.
[13] R. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health using social media text," in *Proc. NAACL- HLT*, 2017.
[14] A. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014.
[15] Apify, "Web scraping and automation platform," [Online]. Available: apify.com.
[16] Streamlit Inc., "Streamlit: A framework for building data applications," [Online]. Available: streamlit.io.
[17] World Health Organization, "Suicide worldwide in the 21st century," WHO, 2021.