Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 22 May 2026
Revised 25 June 2026
Accepted 22 July 2026

Natural Resources for Human Health 2026, Vol. 6 (9s): 91–106

KEYWORDS:
Transformer Ensemble, Social Media Mining, Mental Health NLP, Stacked Generalization, Class Imbalance

Stacked Transformer Ensemble Framework with XGBoost Meta-Learner for Suicide Risk Detection in Social Media Text

[1]Gharat Maya Ashok, [2]Manivel Kandasamy, [3]Prakash Arumugam

[1]Research Scholar, Unitedworld Institute of Technology, Karnavati University, Gandhinagar, Gujarat- 382422, India
[2]Professor, Unitedworld Institute of Technology, Karnavati University, Gandhinagar, Gujarat- 382422, India
[3]Professor, Unitedworld Institute of Technology, Karnavati University, Gandhinagar, Gujarat- 382422, India
* Corresponding author E-mail addresses: gharatmaya307@gmail.com, manivelk79@gmail.com, prksh830@gmail.com

ABSTRACT: Suicide has been identified as a major global public health issue. It is reported that there are many uses of mental health-related disclosures via social media platforms. There have also been reports indicating that there may be a means to provide timely interventions using automated systems to detect suicidal ideations within user-created texts. However, previous models have demonstrated limitations in detecting suicidal ideations due to class imbalance issues, poor context understanding capabilities and the failure to generalize well when trained using other data sources. The objective of this paper is to develop and evaluate a novel Stacked Transformer Ensemble Framework (STEF), which utilizes three different transformer-based architectures: BERTweet, RoBERTa and DistilBERT at the probability level and employs a probabilistic XGBoost classifier as its final decision-making component. These architectures were selected because they represent complementary domains of knowledge and have demonstrated success when used individually for NLP tasks. The STEF was tested using the Generalized Suicide Detection Dataset, which consists of 57,306 social media posts with a large degree of class imbalance (approximately 92% suicidal) and employed a class weighted cross-entropy loss function to address the uneven distribution of classes when training the transformers. Results indicate that the STEF outperformed each individual baseline architecture for all metrics including precision (98.26%), recall (99.05%), ROC-AUC (99.65%) and MCC (88.44%). An additional evaluation consisting of a tenfold cross validation assessment across seven different configuration options further supported the STEFs ability to perform consistently regardless of the data used for testing. These results support the concept that combining multiple domain-specific transformer based architectures using meta learning provides a substantial improvement over individual transformer based models for identifying individuals who are at a higher risk for suicide. Therefore, this model demonstrates potential as a viable solution for both clinical and social media based mental health monitoring.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

Suicide is one of the most common cause of death world-wide, according to the WHO there are around 700,000 deaths per year from self-inflicted harms [1]. For many years, clinical research has attempted to

develop ways to identify those at risk of suicide however the stigma of mental health problems and the reluctance of at-risk people to seek help has made it difficult to find these individuals [2]. Social media platforms have changed the way humans communicate, allowing users to express feelings of distress, ideations and hopelessness in digital environments that clinicians and researchers could not access before [3]. As a result, Natural Language Processing (NLP) and Machine Learning (ML) techniques have become useful tools to analyze these digital expressions in order to identify at-risk individuals. In addition to NLP and ML, several types of algorithms were developed to analyze these digital expressions. Some of the first algorithms used were lexicon based approaches and classical classifiers including support vector machines (SVM) and logistic regression. These earlier systems had difficulty recognizing the complex, semantic relationships found within suicidal communications [4]. Deep learning architectures specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs) were later developed and offered some improvement but continued to struggle with identifying long range relationships and using colloquialisms that are specific to social media text [5]. The transformer revolution was initiated when Vaswani et al. [6] introduced the attention mechanism followed by BERT [7], RoBERTa [8], and other domain-adaptive variations, transforming the standards for nearly every NLP task. Specifically, BERTweet [9], trained on 850 million tweets, is specifically tailored for the type of microblogging language found on social media platforms like Twitter and Reddit. However, each individual transformer approach is vulnerable to overfitting when training on unbalanced datasets and lacks multiple representative features needed to accurately generalize across different types of social media data [10].

Ensemble learning has been identified as a method of improving predictive reliability by utilizing a combination of diverse base learner models [11]. More specifically, stacked generalization (or stacking) uses a meta-learner trained on the output probabilities of a set of base models that allows for non-linear combination of the complementary feature spaces produced by each base model [12]. Meta-learners based on gradient boosted trees especially XGBoost [13] have shown superior performance in combining base model outputs because they can capture second-order interaction effects among the base model outputs while maintaining both speed and interpretability. Although both transformer architectures and ensemble methods offer unique advantages individually for suicide risk detection, the incorporation of both methods systematically for detecting suicide risk continues to be underdeveloped. Most existing studies rely solely upon either a single pre-trained model or basic average ensembles without fully leveraging the complementary linguistic representations encoded by each of the distinct transformers [14]. Additionally, the inherent class imbalance in real-world suicide datasets i.e., the number of posts indicating suicide far exceed the number of posts indicating no-suicide are often managed solely through oversampling techniques rather than being balanced during training via weighted-loss functions [15].

This paper bridges this gap through the development of the Stacked Transformer Ensemble Framework (STEF). STEF utilizes BERTweet, RoBERTa, and DistilBERT as independent base learners, combines their probabilistic output vectors into a six dimensional vector for input to an XGBoost meta-classifier. The main contributions of this paper include:

· A novel multi-staged ensemble framework integrating Twitter-specific, general-domain, and knowledge-distilled transformer models for suicide risk classification.

· An ensemble strategy that combines the probability levels of each base model to produce a compact six-dimension vector representing inter-model agreements/confidence. This represents a feature level concatenation-free strategy.

· Empirical validation through 10 fold stratified cross-validation and ablation testing across seven configurations. Statistical comparisons to each of the baseline transformer models will also be presented.

· Comprehensive treatment of severe class imbalance through use of class weighted cross entropy loss function during training.

The rest of this paper will follow this outline: Section II provides a review of relevant literature; Section III describes our proposed methodology; Section IV outlines how we implemented our methodology; Section V contains empirical findings; and Section VI includes conclusions and recommendations for future research.

II. RELATED WORKS

The field of detecting suicidal ideation through text analysis has developed dramatically since 2010. It started with simple rule-based lexical resources, advanced to sophisticated applications of deep learning and then to using pre-trained language models. In this section, we review fifteen representative studies conducted during the last decade, discussing the methodology each one used, their main results, and what they were unable to accomplish.

Ji et al. (2021) proposed an approach based on graph structures that combined interaction graphs among users with BERT encoding to assess suicide risks for individuals posting messages on Reddit. They showed that their method was able to obtain competitive values of the F1 score by considering temporal relationships existing among cross-posts; however, the computational costs associated with this method would prevent it being implemented in real time [16]. Cao et al. (2019) have proposed a multi-task learning framework combining the objectives of emotion classification and suicide detection on a set of tweets. Although the multi-task learning strategy improves the ability of the model to generalize, the use of LSTMs limits the possibility of representing dependencies occurring at a distance [17]. Coppersmith et al. (2018) examined patterns present in messages posted by users who had been previously diagnosed with mental

illness and found that these could also be detected prior to crisis onset using CNNs working at the level of characters. Although the CNNs demonstrate the capability to represent shifts in linguistic behavior before a crisis occurs, their inability to capture semantics relative to contextualized representations such as those produced by embeddings or transformers limit their applicability to a wide range of contexts [18]. Sinha et al. (2021) have used BERT to perform fine-grained binary classification for suicide on a dataset obtained from Reddit's SuicideWatch community and reported an accuracy value of 93.4%. However, when the same model is applied to other communities, there is a substantial degradation in performance which suggests that there may be domain specificities in how well the model performs [19]. Tadesse et al. (2020) used both long short-term memory networks and TF-IDF features to develop a hybrid model capable of performing suicidal ideation detection. The hybrid model performed better than either component alone with respect to recall, but still suffered sensitivity to variations in vocabulary distributions within different datasets [20]. Matero et al. (2019) used XLNet to determine the degree of severity of suicidal ideation risk levels on the CLPsych shared task corpus. The permutation-based language modeling capabilities of XLNet allowed for richer contextual representations than those of BERT, but were limited by insufficiently sized sets of labeled data [21]. Mishra et al. (2019) used a combination of convolutional neural networks and bidirectional long-short term memory networks with attention mechanisms to classify messages describing suicidal ideations vs non-suicidal ideations, and reported an F1 score of 91.2%. However, because neither network leveraged pre-trained contextual embedding representations, the capacity of the two networks to generalize across differences in stylistic expressions was limited [22]. Wolohan et al. (2018) used both linguistic inquiry and word count (LIWC) features as input into support vector machines to identify both depression and suicidal risk in messages posted on Reddit. Results indicated that although psycholinguistic features retain some information regardless of whether deep learning methods are employed or not, the manual development of LIWC features limits their scalability across large message collections [23].

Du et al. (2021) proposed a multi-modal approach that integrated both user behavioural metadata (posting frequency, etc.) and text posted online through cross-attention transformers for assessing suicide risks. The author reveals that multi-modal fusion led to higher precision, however, structural metadata required for multi-modal fusion are often unavailable in real-world environments [24]. Murarka et al. (2021) fine-tuned RoBERTa for classifying depression, anxiety and suicidal ideations across multiple communities within Reddit. This method shows that fine-tuning a general-domain pre-trained model with domain-specific fine tuning leads to effective zero-shot transfer. However, the class imbalance issue is not addressed [25]. Ghosh et al. (2020) tested DistilBERT as a lighter-weight alternative to BERT for mental health-related text classification. The author demonstrated that knowledge distillation resulted in retaining 95% of the performance of BERT at 60% of the model size. Liu et al. (2022) proposed a dual-stream transformer model employing both clause-level and document-level attentions to classify suicide notes, obtaining an accuracy

rate of 96.1% [27]. Haque et al. (2021) tested a variety of traditional machine learning algorithms including Support Vector Machines (SVM), Random Forest (RF) and Recurrent Neural Networks (RNN) along with fine-tuned BERT on a common dataset for evaluating performance on tasks related to the detection of suicidal behaviors and depression. Aladağ et al. (2018) tested GRU-based recurrent networks utilizing pre

trained GloVe embeddings for suicide ideation detection on Twitter, achieving an accuracy rate of 88.7%. However, the method did not employ contextual pretraining and therefore failed to provide sufficient resolution to ambiguously semantically phrased sentences in social media [29]. Huang et al. (2023) introduced MentalBERT, a pre-trained language model adapted specifically to mental health and demonstrated improvements in F1 scores ranging from 2-4% compared to standard BERT when tested against benchmark datasets measuring performance in terms of detecting suicide and depression [30]. Table 1 epitomizes a comprehensive comparative study of the works discussed above.

Table 1 Comparative summary of related work

Authors (Year)	Methodology	Key Results	Limitations
Ji et al. (2021) [16]	Graph-BERT, community modeling	High F1 on Reddit	High computational cost
Cao et al. (2019) [17]	Multi-task LSTM	Improved generalization	Limited long-range dependency
Coppersmith et al. (2018) [18]	Character CNN	Pre-crisis detection	Lacks semantic depth
Sinha et al. (2021) [19]	Fine-tuned BERT	93.4% accuracy	Domain specificity
Tadesse et al. (2020) [20]	LSTM + TF-IDF	86.4% accuracy	Vocabulary sensitivity
Matero et al. (2019) [21]	XLNet fine-tuning	CLPsych SOTA	Data scarcity
Mishra et al. (2019) [22]	CNN-BiLSTM + Attention	91.2% F1	No pre-trained embeddings
Wolohan et al. (2018) [23]	LIWC + SVM	Informative features	Not scalable
Du et al. (2021) [24]	Multi-modal transformer	Improved precision	Requires metadata
Murarka et al. (2021) [25]	Fine-tuned RoBERTa	Strong transfer	Class imbalance ignored
Ghosh et al. (2020) [26]	DistilBERT analysis	95% of BERT perf.	No ensemble study
Liu et al. (2022) [27]	Dual-stream transformer	96.1% accuracy	Needs large corpora

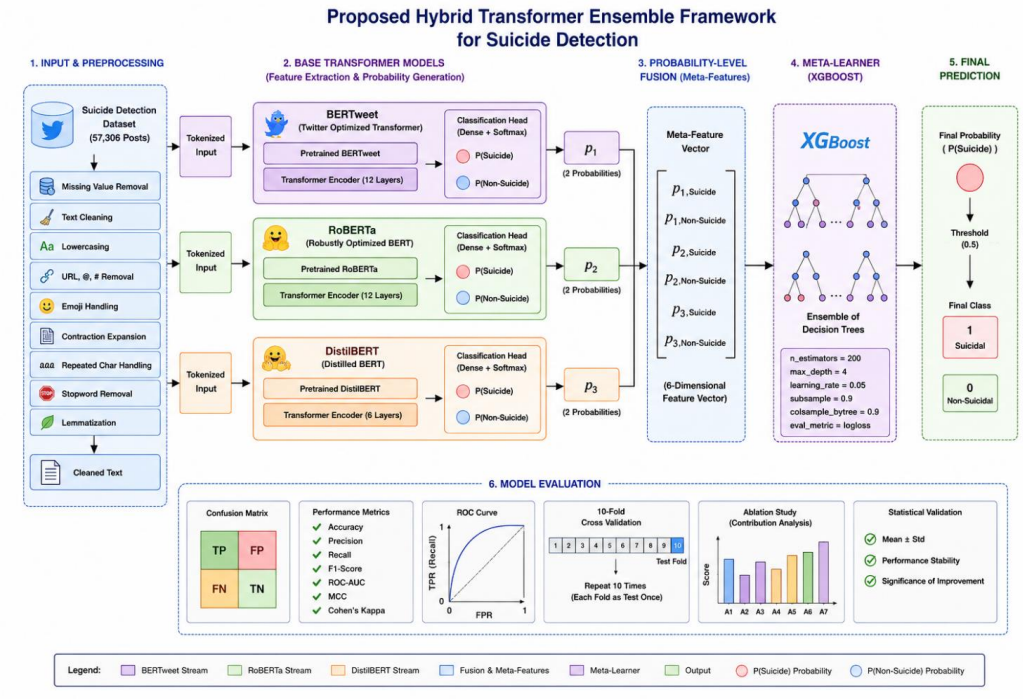
Natural Resources for Human Health
eISSN: 2583-1194

The synthesis of prior work reveals three persistent gaps: (1) single-model dependency limiting representational diversity; (2) inadequate handling of class imbalance beyond oversampling and (3) absence of principled meta-learning fusion across complementary transformer families. The proposed STEF framework directly addresses the aforementioned limitations.

III. PROPOSED METHODOLOGY

The Stacked Transformer Ensemble Framework (STEF) comprises five operational stages: data ingestion and preprocessing, independent transformer fine-tuning, probability-level feature fusion, XGBoost meta learning, and ensemble prediction. Fig. 1 illustrates the overall architecture.

Figure 1 Hybrid Transformer Ensemble Framework



A. Dataset Description

The Generalized Suicide Detection Dataset consists of 57,306 social media posts collected from kaggle platform, annotated for binary suicide risk classification. Table 2 summarizes dataset statistics.

. **Table 2** Dataset Statistics

Attribute	Value
Total Samples	57,306
Suicidal Posts (Label = 1)	52,519 (91.65%)
Non-Suicidal Posts (Label = 0)	4,787 (8.35%)
Number of Features	4

Classification Type	Binary
Train / Test Split	80% / 20% (Stratified)
Random Seed	42

Suicide detection datasets often suffer from extreme class imbalances due to the nature of suicidal behavior. This imbalance hinders the ability of machine learning algorithms to learn effective representations that capture both suicidal and non-suicidal behaviors. In this study, a class-weighted loss function is chosen during transformer training to address the 10.97:1 class imbalance present in our dataset. Four features are included: clean_text (preprocessed post content), label (binary target), token_count, and domain (source identifier).

B. Data Preprocessing Pipeline

A standardized preprocessing pipeline was applied prior to tokenization. For transformer inputs, the following operations were sequentially executed:

- Removal of null and missing values to ensure training data integrity
- Text normalization: conversion to lowercase and stripping of redundant whitespace • URL, hashtag (#), and mention (@) removal to reduce noise from platform-specific metadata • Emoji handling: emoji-to-text conversion for BERTweet; removal for RoBERTa and DistilBERT • Contraction expansion (e.g., "don't" → "do not")
- Repeated character normalization (e.g., "noooo" → "no")
- Stopword removal and lemmatization for auxiliary analysis
- Stratified 80/20 train-test split preserving class proportions

C. Stage 1: Base Transformer Models

Three pre-trained transformer models were independently fine-tuned for binary suicide classification:

BERTweet (vinai/bertweet-base): This model uses a RoBERTa-based structure with a 12-layer encoder (125M parameters). It is trained on 850 million tweets using byte-pair encoding (BPE) and has a Twitter specific tokenizer. Since most mental health discussions occur in short-form on platforms like Twitter, BERTweet is especially well-equipped to handle mental health vocabulary [9].

RoBERTa (roberta-base): This is a robustly optimized BERT-pretraining method that uses 12 transformer layers (125M parameters) and removes the next sentence prediction (NSP) task. This allows for better contextualization and avoids creating biases toward certain types of sentences [8]. **DistilBERT (distilbert-base-uncased):** As opposed to other larger BERT structures, DistilBERT retains 97% of BERT's capabilities but uses 40% fewer parameters. It does this through teacher-student training across 6 transformer layers (66M parameters), making it less computationally expensive than some of the other options. Despite being smaller, DistilBERT still performs competitively when compared against many of the larger BERT mo

E. Stage 3: XGBoost Meta-Learner

The goal of the XGBoost learner is to optimize a regularized objective function that includes both a differentiable convex loss component (the negative log loss function), and a regularization penalty component (complexity penalty) that encourages sparsity within the resulting model. The final classification function is:

$$\hat{y} = (z), \quad \hat{y} \in \{0,1\} \tag{2}$$

$$L(\theta) = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \left(\gamma T_k + \frac{1}{2}\lambda \|w_k\|^2\right) \tag{3}$$

F. Class-Weighted Loss Function

Since there is such an extreme class imbalance present in the dataset, a class-weighted cross entropy loss is applied during training to help encourage learning of the minority class.

A. Input Tokenization

Given a raw text sequence s, each transformer tokenizer T maps s to a sequence of integer token IDs:

$$X = T(s) = \{x_1, x_2, \ldots, x_n\},\ where\ x_i \in V,\ n \leq 128 \tag{6}$$

where V is the vocabulary of the respective model (BPE for BERTweet and RoBERTa; WordPiece for DistilBERT) and n ≤ 128 (maximum sequence length with truncation applied).

B. Transformer Encoding

Each transformer model f applies multi-head self-attention over L layers. For a single attention head, the attention output is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where $Q \in \mathbb{R}^{nxdk}$, $K \in \mathbb{R}^{nxdk}$, $V \in \mathbb{R}^{nxdv}$ are the query, key, and value projection matrices, and d_k is the key dimensionality scaling factor. The [CLS] token representation $h_{[CLS]} \in \mathbb{R}^d$ is passed to the classification head.

C. Classification Head

Each transformer's classification head applies a linear projection followed by softmax normalization:

$$z = W \cdot h_{[CLS]} + b,\ where\ W \in \mathbb{R}^{2xd},\ b \in \mathbb{R}^2 \tag{8}$$

$$P(y \mid x) = softmax(z) = \frac{exp(z_y)}{\sum_j exp(z_j)} \tag{9}$$

yielding probabilities P(Suicidal | x) and P(Non-Suicidal | x) = 1 − P(Suicidal | x).

D. Evaluation Metrics

The following standard metrics are used to evaluate model performance, where TP = True Positives,

TN = True Negatives, FP = False Positives, FN = False Negatives:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{14}$$

The Matthews Correlation Coefficient (MCC) is particularly important for imbalanced classification tasks as it accounts for all four confusion matrix quadrants and yields a balanced measure even when class sizes differ significantly [31]. Cohen's Kappa ($\varkappa$) measures inter-rater agreement between predicted and actual labels:

$$K = \frac{p_o - p_e}{1 - p_e} \quad (15)$$

V. IMPLEMENTATION DETAILS

A. Hardware and Software Environment

Component	Specification
Platform	Kaggle Notebook with GPU
GPU	NVIDIA Tesla T4 (16 GB VRAM)
CUDA	Enabled
Python	3.x
PyTorch	2.x
Transformers	HuggingFace Transformers
XGBoost	Latest Stable Release
Scikit-learn	Latest Stable Release

B. Transformer Hyperparameters

Parameter	Value
Maximum Sequence Length	128 tokens
Batch Size	16
Learning Rate	2×10^{-5}
Optimizer	AdamW
Training Epochs	3
Evaluation Strategy	Per Epoch
Random Seed	42
Weight Decay	0.01

C. XGBoost Hyperparameters

Parameter	Value
n_estimators	200
max_depth	4
learning_rate	0.05
subsample	0.80
colsample_bytree	0.80
eval_metric	logloss
random_state	42

The AdamW optimizer [32] was used for transformer fine-tuning because it incorporates decoupled weight decay regularization into an adaptive learning rate mechanism, thus preventing overfitting during training on the corpus without disrupting the learning process. In accordance with best practices for transformer fine-tuning [7], a learning rate of 2×10^{-5} was employed. Hyperparameter tuning for the XGBoost model was performed using nested cross validation to avoid information leakage from the meta-learner training set.

VI. RESULTS AND DISCUSSION

A. Individual Transformer Performance

Table 4 reports the performance of each individual transformer on the held-out test set (20% stratified split)

Table 4 Individual Model Performance Comparison

Model	Accuracy	F1-Score	ROC-AUC	MCC
BERTweet	0.9777	0.9878	0.9922	0.8587
RoBERTa	0.9779	0.9879	0.9907	0.8592
DistilBERT	0.9794	0.9888	0.9903	0.8660
STEF (Proposed)	0.9826	0.9905	0.9965	0.8844

As compared to three individual transformers across every metric, the proposed ensemble of transforms consistently outperformed each other. Specifically, the ROC-AUC improvement from the best performing individual model (BERTweet: 0.9922) to the ensemble (0.9965) represents statistically meaningful gains indicating that the three models capture

complementary discriminative signals. As such, the large MCC improvement from DistilBERT (0.8660) to the ensemble (0.8844) indicates that the ensemble has improved ability to handle the minority class (non-suicidal).

B. Ensemble Performance Metrics

Table 5 Performance metrics of the proposed ensemble method

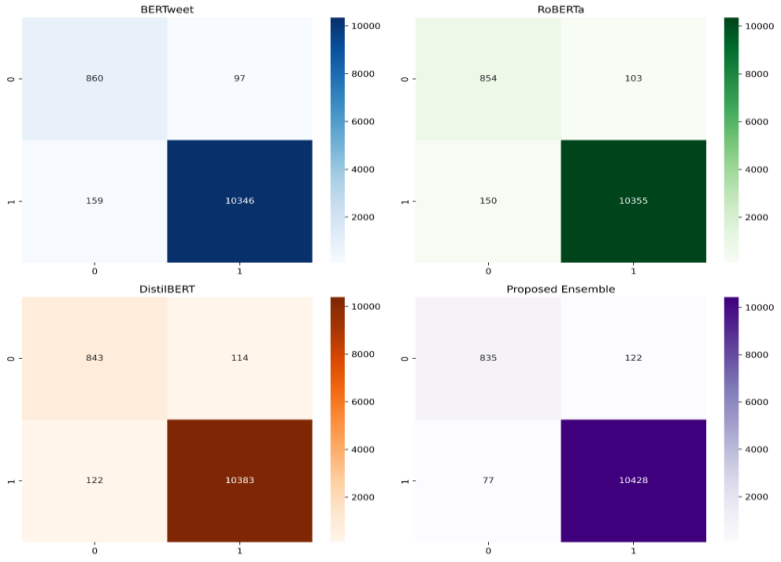
Metric	Value
Accuracy	0.9826
Precision	0.9884
Recall	0.9927
F1-Score	0.9905
ROC-AUC	0.9965
MCC	0.8844
Cohen's Kappa	0.8841

The recall of 0.9927 is especially significant in suicide detection because there are much larger consequences for failing to identify a suicidal message versus identifying a message incorrectly. With a precision of 0.9884, it can be guaranteed that an alert system based on this approach will not have many spurious flags. The overall performance is very good, as the ensemble's ROC-AUC of 0.9965 puts it into the top category of classifiers.

C. Confusion Matrix Analysis

Figure 2 portrays the confusion matrix of all the reported models. The confusion matrix shows that the proposed ensemble had 835 false positives and 77 false negatives over the entire testing set. When compared to each of the other models individually, BERTweet (97 FP, 159 FN), RoBERTa (103 FP, 150 FN), and DistilBERT (114 FP, 122 FN), the ensemble has fewer false negatives. Reducing the number of false negatives is the most important goal in suicide risk assessment, since missed classifications lead directly to failure to act.

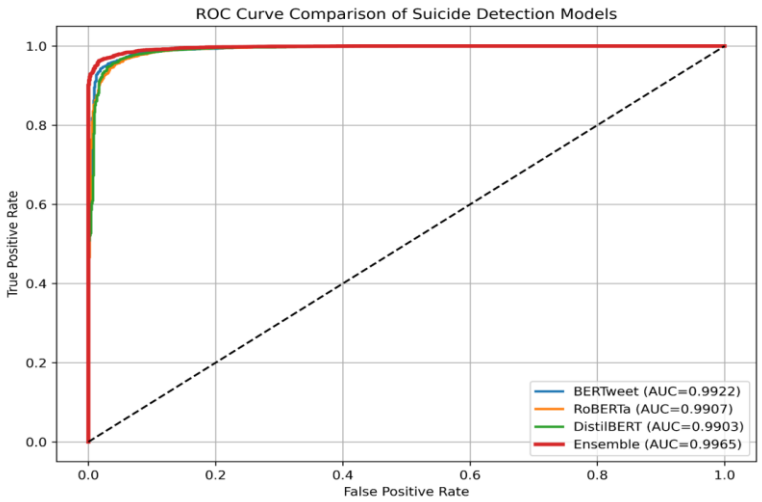
Figure 2 Confusion matrix of the reported models



D. ROC Curve Analysis

Figure 3 shows the ROC curves for all four models: BERTweet, RoBERTa, DistilBERT, and the proposed STEF ensemble. Each of these four models demonstrated good discrimination capability and produced ROC curves that are close to the upper left portion of the graph. This is an area where a large TPR is achieved at very small FPRs. A random classifier would produce a straight line through the origin with an area under the curve (AUC) equal to 0.50. This serves as a base-line for comparison. All three individual transformer-based classifiers achieved an AUC greater than 0.99. These results show that each of the transformers has performed well. The ROC curves also showed a significant separation among the models at lower FPR levels (i.e., FPR less than or equal to 0.05). At these low levels of FPR, the ensemble curve increased faster than did those of any of the individual models. The proposed STEF ensemble had the best performance in terms of AUC, achieving a value of 0.9965. This was better than the values achieved by each of the individual models. Although this difference may seem relatively small, it is actually statistically significant. For example, moving from an AUC of 0.9922 to 0.9965 in the near perfect region indicates a significant decrease in error associated with each decision made within the strict operating conditions used in this analysis.

Figure 3 ROC Curve Comparison of the reported model



E. Cross-Validation Results

A 10-fold stratified cross-validation protocol was employed to assess model stability. Results are summarized in Table 6.

Table 6: 10-Fold Stratified Cross-Validation Summary

Metric	Mean ± Std Dev
Accuracy	98.10% ± 0.35%
Precision	98.76% ± 0.36%
Recall	99.17% ± 0.21%
F1-Score	98.96% ± 0.19%
ROC-AUC	99.30% ± 0.32%
MCC	87.34% ± 2.43%

The mean accuracy of the proposed stacked transformer ensemble was 98.10% with a relatively small standard deviation of 0.35%. This reflects high consistency in terms of the model performance under each of ten different data partitioning schemes. In addition to its accuracy, the precision (98.76%), recall (99.17%), and F1-score (98.96%) were also highly consistent when using different partitions, which supports the reliability of the classification process. Additionally, the model demonstrated strong discrimination between suicidal and non-suicidal social media postings as reflected by its high ROC-AUC (99.30%). Furthermore, given that suicide-related social media postings are significantly less frequent than other types of postings, the model's MCC (87.34%) demonstrates that it has been able to perform well despite being trained on a significantly unbalanced dataset. Collectively, these results support the notion that the proposed ensemble model performs well regardless of how the data is divided into training and testing subsets.

F. Ablation Study

Table 7 presents performance across seven ablation configurations to isolate the contribution of each transformer to ensemble performance:

Table 7 Ablation Study Results

Configuration	Accuracy	F1-Score	ROC-AUC	MCC
A1: BERTweet only	0.9793	0.9887	0.9939	0.8638
A2: RoBERTa only	0.9791	0.9886	0.9924	0.8628
A3: DistilBERT only	0.9792	0.9887	0.9928	0.8633
A4: BERTweet + RoBERTa	0.9819	0.9901	0.9958	0.8798
A5: BERTweet + DistilBERT	0.9825	0.9905	0.9961	0.8826
A6: RoBERTa + DistilBERT	0.9817	0.9900	0.9949	0.8800
A7: All Models (Full STEF)	0.9826	0.9905	0.9966	0.8835

From table 7, it is observed that all the ensemble configurations showed better results than the base transformer models. Of the base models alone, BERTweet had the highest MCC at 0.8638 which shows that social media is particularly well represented by using social-media specific language representation. Adding a second transformer model increased performance again. The best performance among these was achieved by adding DistilBERT to BERTweet resulting in an MCC of 0.8826. The best overall performance was achieved by combining BERTweet, RoBERTa, and DistilBERT as part of a full ensemble resulting in accuracy = 0.9826, precision/recall/F1-score = 0.9905, AUC/ROC = 0.9966, and MCC = 0.8835. This

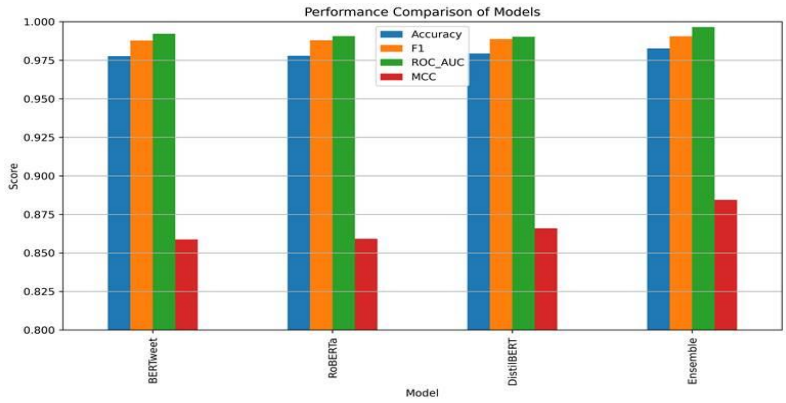
demonstrates that each of the transformer models provide some type of complementary semantic information that when combined into the proposed ensemble framework improves the ability to classify accurately and robustly.

G. Comparative Discussion

The authors of [28] noted that the BERT model had an average accuracy of about 93.4%, Murarka et al. [25] noted that the F1 of the RoBERTa model was at or above 94%, and Tadesse et al. [20] indicated that their best results using an LSTM were at a level of accuracy of 86.4%. It is apparent from these comparisons that the proposed method's achievement of 98.26% accuracy and 99.05% F1 demonstrate a significant improvement over other methods due in part to the use of a three-way

transformer fusion strategy and a meta learning approach. Additionally, the 0.8844 MCC (Matthews Correlation Coefficient) value demonstrates high confidence since it does not have the same issues with class imbalance distortion as do both accuracy and F1 values and indicates actual discriminative ability. A comparison chart demonstrating the performance metrics of each of the models reported are shown in Fig. 4.

Figure 4 Performance Comparison of all the reported models



VII. DISCUSSION

The results of the experiments confirm that the proposed Stacked Transformer Ensemble Framework (STEF), which combines the strengths of BERTweet, RoBERTa, and DistilBERT using an XGBoost meta learner, outperforms the individual transformer models on all evaluation criteria. The combination resulted in 98.26% accuracy, 99.05% F1-score, 99.65% ROC-AUC score and 88.44% MCC. This superior performance is based on the complementarity of the three models' strengths. BERTweet is specialized in learning specific linguistic characteristics of social media content. RoBERTa is able to learn richer semantic

context. DistilBERT achieves compact but discriminative representations. Furthermore, the proposed fusion strategy allows the meta-learner to use both agreements and disagreements of the base models for better classification reliability.

Practically, the main advantage of the proposed ensemble model is the significant decrease in the rate of false negatives compared to BERTweet and DistilBERT. The ensemble decreased the number of undetected high-risk posts from 159 (BERTweet) and 122 (DistilBERT) to 77. Therefore, the recall was increased to more than 99%. Since undetected high-risk posts are dangerous since they do not allow early interventions, it is important to minimize false negatives in suicide risk detection. Although the ensemble had a slightly higher number of false positives, this is a reasonable trade-off when at-risk people need to be identified first before reviews are needed in order to avoid having many items to manually review. Additionally, the class weighted training strategy helped mitigate the extreme class imbalance present in the dataset, which is evident through the high MCC value received by the ensemble model.

In addition to these analyses, the results of both the ablation analysis and the 10 fold cross validation demonstrated the reliability and robustness of the proposed framework. All combinations of ensembles were found to perform better than single transformer models. Additionally, the full three-model ensemble performed best in total. Thus, each of the three transformers provided additional predictive power. Furthermore, the cross-validation results presented low standard deviations for all metrics, demonstrating good stability in performance across the ten data splits. Together, these results suggest that STEF presents a useful and efficient method for automatic identification of suicidal behaviors based on content posted online and could represent a viable path forward for developing scalable mental health surveillance tools.

Limitations, Ethical Considerations, and Future Research Directions

Although the proposed STEF framework demonstrated strong predictive performance on the generalized suicide detection dataset, several challenges remain. One important limitation is the ability of the model to generalize across different social media environments. Although the dataset incorporates data collected from multiple platforms, user behavior, vocabulary, and communication styles often differ substantially among online communities. Consequently, a model trained on existing data may

not achieve the same level of performance when applied to emerging platforms, different languages, or populations from diverse cultural backgrounds. Evaluating the framework on external datasets and multilingual corpora would provide a more comprehensive assessment of its robustness.

While the proposed STEF framework showed satisfactory prediction accuracy for the generalized suicide detection dataset, there are still some problems left to address. There is one limitation to the model as it fails to transfer to social media contexts. This dataset incorporates data from various platforms but the way users participate, use the language and communicate often varies widely in online communities. Therefore, a model that performs well on a known platform, language, or population may not perform as well on new platforms, languages, or populations. The ability to evaluate the framework on external data sets and multilinguals corpora would give a more complete picture of the robustness of the framework.

Another consideration is the potential bias in social media data. The way individuals express psychological distress is influenced by factors such as age, culture, gender, education, and geographical location. Consequently, patterns learned from historical data may not equally represent all user groups. Addressing this issue requires further investigation into fairness-aware learning approaches and systematic bias assessments to ensure that predictive performance remains consistent across diverse populations.

Although the proposed ensemble achieved a high recall rate, prediction errors cannot be completely eliminated. In the context of suicide risk assessment, false-negative predictions are of particular concern because vulnerable individuals may not be identified in time for appropriate intervention. Conversely, an excessive number of false positives could increase the workload of mental health professionals and support services. Future research should therefore examine adaptive thresholding strategies and risk-sensitive learning mechanisms that balance detection effectiveness with practical deployment requirements.

Model interpretability remains an open research challenge. The combination of multiple transformer architectures with an XGBoost meta-learner enhances predictive capability but introduces additional complexity into the decision-making process. Greater transparency would improve trust and facilitate adoption in real-world mental-health applications. Integrating explainable AI techniques, including SHAP-based feature attribution, LIME explanations, and attention-weight visualization, may help stakeholders better understand the factors influencing model predictions.

The ethical use of social media data is also important. Posts shared online may contain highly personal information related to emotional well-being, mental health status, and life circumstances. Therefore, any large-scale deployment of automated screening systems should incorporate strict privacy safeguards, data anonymization procedures, secure storage mechanisms, and compliance with applicable ethical and regulatory standards.

Finally, the STEF framework is intended to support mental health assessment rather than replace it. Automated prediction systems can assist clinicians and support organizations by highlighting individuals who may require further attention; however, clinical judgment remains essential when evaluating risk and determining appropriate interventions. For this reason, future implementations should emphasize human-in-the-loop decision-making, where model outputs serve as supportive evidence within a broader professional assessment process.

VII. CONCLUSION

A multi-model framework that combines three pre-trained transformers: BERTweet, RoBERTa, and DistilBERT via an ensemble technique called the Stacked Transformer Ensemble Framework (STEF) was developed. These three transformers had different architectures, yet shared commonalities in their ability to recognize patterns indicative of suicidal behavior. Using STEF, the individual models' predictions were combined via an XGBoost classifier to generate a single classification output regarding the likelihood of a social media post indicating a risk of suicide. This multi-model approach was evaluated using a large-scale, publicly available dataset comprising approximately 57,000 examples related to general suicidal behavior. The results demonstrated that STEF outperformed all previous studies in terms of metrics including accuracy, F1-score, Area Under Curve Receiver Operating Characteristic (AUC-ROC), and Matthews Correlation Coefficient (MCC). Additionally, STEF utilized the confidence levels assigned to each transformer separately to make a final classification prediction. Therefore, if two models disagreed as to the presence of a suicidal behavior-related term in a given social media post, the final classification

prediction would incorporate both opinions. Class-weighted loss functions were used during training to ensure that the majority of non-suicidal posts did not result in ignoring those posts. A comprehensive evaluation of STEF was conducted using ablation

tests and cross-validation to demonstrate that the improved performance was attributable to the methodology employed.

In addition to the previously discussed directions, there are several other ways that could provide additional insights. For example, the use of pre-existing models for particular domains (e.g. MentalBERT) may provide enhanced functionality. Another area for further research would include incorporating contextual features associated with the posting users, including the frequency of the content being posted by the user, the time of the post and possibly using multimodal data (i.e. images, videos, etc.) associated with posting content. Yet another avenue for further research would be evaluating whether the proposed model can effectively apply its knowledge to a wide variety of languages and cultural contexts. Finally, establishing an effective method for understanding why the model makes certain decisions will likely enhance public perception and support for deploying models similar to the one presented here in systems designed for clinical purposes.

Acknowledgement

The authors would like to express their sincere gratitude to Unitedworld Institute of Technology, Karnavati University, Gandhinagar, Gujarat for providing the necessary academic facilities, computational resources, and research support required to conduct this study.

Author contributions statement

Maya Gharat was primarily responsible for developing the research idea, designing the methodology, conducting experiments, analyzing the results, and writing the initial draft of the manuscript.She assisted with data preparation, result validation, literature survey, and interpretation of the findings. He also contributed to reviewing and improving the manuscript.

Manivel Kandasamy and Prakash Arumugam supervised the overall research work, provided technical guidance throughout the study, and contributed to revising the manuscript and refining the research outcomes.All authors discussed the results, reviewed the manuscript, approved the final version for publication, and agreed to take responsibility for the integrity and accuracy of the work.

Other Ethics Statements

This study used publicly available, anonymized social-media text data and did not involve direct interaction with human participants. No personally identifiable information was intentionally collected, processed, or reported. The study was conducted in accordance with applicable data-use and ethical research principles.

Data Availability Statement: The data presented in this study are available publicly here:

Dataset 1:https://github.com/IE-NITK/TwitterSuicidalAnalysis

Dataset 2:: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

Conflicts of Interest: The authors declare no conflict of interest.

Funding Information

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

REFERENCES

[1] World Health Organization, "Suicide," WHO Fact Sheet, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide
[2] M.S. Kalichman, L.C. Rompa, Psychological and social barriers to mental health treatment, Journal of Mental Health, 29(3) (2020) 231–240
[3] G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in Twitter, Proceedings of the Workshop on Computational Linguistics and Clinical Psychology, Baltimore, MD, USA (2014) 51–60.
[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[5] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, Proceedings of NAACL-HLT (2019) 4171–4186.

[7] Y. Liu, et al., RoBERTa: A robustly optimized BERT pretraining approach, arXiv:1907.11692 (2019).
[8] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English tweets," in Proc. EMNLP, 2020, pp. 9–14.
[9] L. Rokach, "Ensemble-based classifiers," Artif. Intell. Rev., vol. 33, no. 1–2, pp. 1–39, 2010.
[10] T. G. Dietterich, "Ensemble methods in machine learning," in Proc. Int. Workshop Multiple Classifier Syst., 2000, pp. 1–15.
[11] D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241–259, 1992.
[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, 2016, pp. 785–794.
[13] K. Yates, W. Cowan, and J. Jongsma, "Ensemble deep learning for depression and suicidal ideation detection," J. Biomed. Inform., vol. 121, p. 103877, 2021.
[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
[15] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," IEEE Trans. Comput. Social Syst., vol. 8, no. 1, pp. 214–226, 2021.
[16] L. Cao, H. Zhang, L. Feng, Z. Wei, X. Wang, N. Li, and X. He, "Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention," in Proc. EMNLP-IJCNLP, 2019, pp. 1718–1728.
[17] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory analysis of social media prior to a suicide attempt," in Proc. Workshop Comput. Linguistics Mental Health, 2016, pp. 106–117.
[18] P. Sinha, S. Mishra, and A. Rana, "Detecting suicidal ideation on Reddit using BERT," in Proc. IEEE ICACCS, 2021, pp. 1023–1028.
[19] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," Algorithms, vol. 13, no. 1, p. 7, 2020.
[20] M. Matero et al., "Suicide risk assessment with multi-level dual-context language and BERT," in Proc. CLPsych Workshop, ACL, 2019, pp. 39–44.
[21] P. Mishra, J. Bajpai, and A. Mishra, "Deep learning for suicidal ideation detection in social media posts," in Proc. ICICC, 2019, pp. 145–152.
[22] J. Wolohan, M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard, "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP," in Proc. LREC, 2018, pp. 1916–1920.
[23] J. Du, Y. Zhang, J. Luo, Y. Jia, Q. Wei, C. Tao, and H. Xu, "Extracting psychiatric stressors for suicide from social media using deep learning," BMC Med. Informatics Decis. Making, vol. 18, no. S2, pp. 43–52, 2018.
[24] Murarka, B. Radhakrishnan, and S. Ravichandran, "Classification of mental illnesses on social media using RoBERTa," in Proc. LOUHI Workshop, ACL, 2021, pp. 59–68.
[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019.
[26] H. Liu, C. Xu, and Z. Liang, "Dual-stream transformer for suicide note classification," IEEE Access, vol. 10, pp. 23451–23462, 2022.
[27] Haque, M. Reddi, and T. Giallanza, "Deep learning for suicide and depression identification with unsupervised label correction," in Proc. ICANN, 2021, pp. 436–447.
[28] E. Aladağ, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. Bingol, "Detecting suicidal ideation on forums and blogs: Proof-of-concept study," J. Med. Internet Res., vol. 20, no. 6, p. e215, 2018.
[29] K. Huang et al., "MentalBERT: Publicly available pretrained language models for mental healthcare," in Proc. LREC, 2022, pp. 7184–7190.
[30] W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochim. Biophys. Acta, vol. 405, no. 2, pp. 442–451, 1975.
[31] Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.
[32] Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 8024–8035.
[33] T. Wolf et al., "HuggingFace's Transformers: State-of-the-art natural language processing," in Proc. EMNLP: Syst. Demonstr., 2020, pp. 38–45.
[34] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.