scientific reports

OPEN

Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique

Daniyal Alghazzawi[1], Hayat Ullah[2], Naila Tabassum[2], Sahar K. Badri[1] & Muhammad Zubair Asghar[2]✉

This research presents a novel framework for distinguishing between actual and non-suicidal ideation in social media interactions using an ensemble technique. The prompt identification of sentiments on social networking platforms is crucial for timely intervention serving as a key tactic in suicide prevention efforts. However, conventional AI models often mask their decision-making processes primarily designed for classification purposes. Our methodology, along with an updated ensemble method, bridges the gap between Explainable AI and leverages a variety of machine learning algorithms to improve predictive accuracy. By leveraging Explainable AI's interpretability to analyze the features, the model elucidates the reasoning behind its classifications leading to a comprehension of hidden patterns associated with suicidal ideations. Our system is compared to cutting-edge methods on several social media datasets using experimental evaluations, demonstrating that it is superior, since it detects suicidal content more accurately than others. Consequently, this study presents a more reliable and interpretable strategy (F1-score for suicidal = 95.5% and Non-Suicidal = 99%), for monitoring and intervening in suicide-related online discussions.

Keywords Machine learning, Suicidal ideation, Social media text, Ensemble models, Explainable AI

In today's era social media platforms have transformed into spaces for people to share their emotions, ideas and experiences. Among these expressions discussions, about suicidal thoughts pose a significant challenge and an opportunity, for early help and intervention. Differentiating between suicidal feelings and those that may not indicate a real danger of self- harm is crucial[1].

In today's era social media platforms have become treasure troves of thoughts and feelings including discussions, about mental health challenges and thoughts of suicide. However, the abundance and intricate nature of these written expressions Present obstacles for health professionals and support networks trying to recognize and aid individuals in need. Even though traditional machine learning systems have their benefits, they often operate like arcane contraptions with little understanding of the decision-making process, which undermines confidence in and understanding of their predictions. Furthermore, in order to reliably identify non-suicidal thoughts and observations in social media writing, a deep understanding of language, context, and human behavior is necessary. Current models consider this task to be very difficult due to the complexity and diversity of human communication. The goal of this research is to control these issues by creating an explainable artificial intelligence (XAI) system that employs integrated techniques to distinguish between suicidal and non-suicidal thoughts expressed in social media posts. The system will utilize an improved approach to merge the advantages of machine learning models, while offering easily understandable rationales, for its forecasts. This strategy not only aims to enhance the accuracy of identifying thoughts but also to make the algorithmic decision making process transparent allowing mental health professionals to understand the key factors influencing the models outcomes. By enhancing AI capabilities in this area the study strives to provide an understandable tool, for early detection and intervention ultimately aiding in more effective mental health assistance and suicide prevention efforts online.

Why is it important to detect suicidal ideation: Suicide is a public health concern influenced by various personal and societal factors such, as trauma, mental and physical health challenges, feelings of isolation,

[1]Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [2]Gomal Research Institute of Computing (GRIC), Faculty of Computing, Gomal University, D. I. Khan (KP), Pakistan. ✉email: mzubairgu@gmail.com

hopelessness and stress. The World Health Organization (WHO) highlighted the seriousness of this issue by reporting than 700,000 suicides in 2019, accounting for 1% of fatalities. This alarming data prompted the WHO to release guidelines to enhance suicide prevention efforts on a scale. It's important to note that suicide attempts are more common than suicides among young people emphasizing the prevalence of suicidal thoughts. These thoughts can range from considerations to planning or unsuccessful actions. Given that Adolescents affected by these issues early detection of suicidal ideation is crucial, in preventing tragedies and preserving lives[2].

Research motivation

Our motivation of this research is driven by the need to bridge the gap between the predictive power of Artificial Intelligence (AI) and the demand for transparency in mental health applications. Our study aims to enhance mental health outcomes by developing interpretable and reliable methods for identifying potential suicidal ideation expressed on social media. By providing explainable AI (XAI) capabilities, the system can empower mental health professionals to better understand the factors influencing model predictions, ultimately facilitating earlier intervention and improving suicide prevention efforts. It's important to differentiate between expressions of suicidal thoughts and non-suicidal mentions when using social media posts to help mental health. This challenging task requires efficient predictive models that can understand the nuanced intricacies of human interaction. Artificial Intelligence (AI), through learning and ensemble techniques has shown great potential in analyzing and understanding large amounts of text data. However, using these technologies in areas like health monitoring and support brings up important ethical concerns, with transparency and explain ability being key issues, and especially, when human lives and mental health are at stake. This is where Explainable AI (XAI) steps in as a link between the abilities of AI and the human need for clear relatable explanations, on how AI systems make their decisions[3].

Combining XAI principles with techniques to differentiate between genuine and non-suicidal thoughts shared on social media is a new and potentially game changing approach. Ensemble methods, which merge predictions from models to enhance accuracy, can greatly benefit from XAIs insights. This not improves performance but also provides insight into the reasoning behind the models decisions. This focus on both performance and transparency is essential, in health applications where misinterpretations can have consequences and the ethical considerations of automated decision making are significant[4].

While AI methods and XAI show promise, in detecting thoughts on social media there is still a lack of research in this area. Not many studies have explored using these technologies in an ethical way.

Research objectives

The objective of the work is to create a binary-label text categorization system that can discriminate between suicidal and non-suicidal thoughts. The goal is to classify each review and assign it a class label $Si \in \{Suicidal, Non - Suicidal\}\ provided\ D = \{d1, d2...dn\}$ as user input. The XAI module will offer easily understandable rationales, for its predictions. This study aims to focus on the following main goals:

1) Various machine learning models to accurately distinguish between suicidalor Non-Suicidal Ideations in social media posts.
2) To combine this approach with Explainable AI techniques ensuring that the decision making processes of the AI system are clear and understandable for humans.
3) Assess the efficiency of the proposed model across measures such as accuracy, precision, recall and interpretability using a selected dataset of social media content labeled for suicidal ideation.
4) To delve into the impacts of this research on health professionals, AI experts and social media platforms with a specific emphasis on ethical considerations the potential for early intervention and fostering mental well being, in online environments.
5) Explore the impact of our discoveries, on health professionals, AI experts and social media platforms by emphasizing the benefits of early intervention and considering the ethical implications of using AI in mental health settings.

Research contributions

1) The proposed study is significant because it employs an ensemble learning technique to classify suicidal thoughts based on user-generated textual content. This study expands on the work of[3], who used an suicidal dataset to develop a model that used a gradient-boosted decision tree and SVM to categorize textual material into suicidal and non-suicidal thoughts types. Our approach, on the other hand, goes beyond, applying several classifiers via ensemble learning to categorize suicidal thoughts into binary classes.
2) The proposed ensemble stacking classifiers investigate to combine predictions from many models. A stacking classifier is a super multi-layer that works as an ensemble learning technique to increase prediction accuracy.
3) This research aims to make contributions, to the fields of AI, mental health and social media analytics. It will introduce an AI based method that combines techniques with XAI for better predictability and interpretability.
4) In terms of accuracy scores, the suggested ensemble system outperforms state-of-the-art systems.
5) Additionally it will offer insights into distinguishing genuine vs non suicidal expressions on social media improving our understanding of online distress communication. Moreover it will set an example for the use of AI in health by emphasizing transparency and explain ability to build trust and ensure responsible deployment. In essence this study presents a morally sound approach to the intersection of technology and mental health. By focusing on AI it seeks to improve the implementation of AI in identifying thoughts on social

platforms. This paves the way, for research and interventions that blend technology with human centered values.

This work’s following sections are organized as follows: A thorough overview of relevant works in the field of suicidal and non-suicidal ideations classification is provided in Section "Related work". The structure and methodology used in our research are described in Section "Proposed methodology". The outcomes of our experiments are shown in Section "Prediction Phase", along with a thorough analysis of the data. Section “Conclusions and future work” explores the implications and constraints of our methodology. Lastly, we conclude the work in Section6 with a summary of the key concepts and suggestions for further research directions.

Related work

Suicidal thoughts recognition-related literature review is presented in this section.

Based on the research, by Tadesse et al.[5] which focused on detecting thoughts of suicide on Reddit using machine learning this study delves into the effectiveness of learning for early detection. We introduce a model that combines LSTM and CNN and assess its performance compared to classification methods. The tests show that this integrated neural network design along with word embedding strategies produces outcomes in categorizing posts potentially serving as a useful tool, for preventing suicides. A model known as the Attention Convolution Long Short Term Memory (ACL) was developed by Chadha and Kaushik[6]. In order to identify indications of thoughts on media, this model integrates CNNs, LSTM, and attention mechanisms. The ACL model outperformed previous techniques by employing grid search to optimize hyper-parameters, particularly when GloVe embedding was used. When using Random embedding, it produced results with up to 88.48% accuracy, 87.36% precision, 90.82% F1 score, 79.23% specificity, and 94.94% recall. Aldhyani et al.[7] used deep learning and machine learning in combination with TF-IDF and Word2Vec text representation analysis on Reddit datasets to develop a strategy for predicting suicidal ideation. CNN + BILSTM and XGBoost have been used in the two studies to accurately recognize suicidal as well as non-suicidal from social media text based on textual and LIWC-22-based variables.

Renjith et al.[8] devised a method to examine social media communications for indications of suicidal ideation by combining CNN, LSTM, and attention mechanisms. Their model outperformed other models throughout testing, achieving an F1 score of 92.6% and an accuracy rate of 90.3%.

Zhang et al.'s work[9] investigated the use of deep learning and transfer learning approaches to find suicide-related stressors on Twitter. They improved their methods by using an existing dataset that had text annotations. Using LSTM, BiTCN, and self attention to identify behaviour in media posts, Choi et al.[10] developed the LSTM Attention BiTCN, a combination model. This model outperformed the baseline with an F1 score of 0.9405 and accuracy. The work conducted by[11] introduces an approach that uses supervised learning to detect suicidal ideation in internet content early on. Through the analysis of user language and discussed themes, it provides insights for an early warning system. It uses statistical, linguistic, and topic aspects to analyze negative emotions, family-related conversations, and social issues. The study compares six classifiers and provides benchmarks for the detection of suicidal ideation on Twitter and Reddit Suicide Watch. Chinese social media users were surveyed by Cheng et al.[12] to determine suicide risk variables, such as stress levels, depression, anxiety, and Weibo communication. Logistic regression and SC-LIWC were used to analyze Weibo postings in order to identify correlations with risk factors. Those who displayed risk characteristics were categorized by an SVM model.

The study conducted by[13] focused on the difficulty of predicting the risk of suicide, which is made harder by a lack of data and the negative view, on health issues. By studying discussions in anonymous Reddit groups scientists create statistical techniques to recognize signs that suggest a move towards thoughts of suicide. These signs help anticipate risks related to thoughts, in the future. The study also examines the moral consequences of this research. The study conducted by[14] delves into the issue of detecting thoughts, on media by creating a unique set of words related to suicide and utilizing machine learning techniques particularly through Weka to study Twitter data gathered via Twitter4J. By using semantic sentiment analysis with WordNet as a basis the research shows that its methodology can effectively identify indicators of ideation in tweets. The success of this method is confirmed by its precision and accuracy, in analyzing sentiments linked to thoughts of suicide. The research conducted by[15] assesses deep learning algorithms, C LSTM for identifying suicidal ideation thoughts. It showcases its effectiveness compared to learning and machine learning models, in classifying text based on the results of experiments. The researchers[16] in their research created a technique to evaluate how the language used in health Reddit communities can influence the likelihood of suicidal thoughts. By utilizing a model that incorporates both evaluations and segmented propensity score analysis they emphasized the importance of self esteem and social network support in reducing this risk. The results emphasize the need for developing resources to strengthen support, within communities. The study[17] introduced SAIPH, an AI system that forecasts thoughts of suicide by examining topics from Twitter data. By merging networks and random forest models it demonstrates precision, in detecting suicidal ideation. Tested on a scale SAIPH shows promise for evaluating suicide risk in settings especially for predicting immediate danger, within a span of 10 days. Recent research highlights the potential of ensemble techniques in boosting performance across various disciplines. Alsulami et al.[18] demonstrated that ensemble methods can enhance student achievement in e-learning environments. Similarly, Moradi et al.[19] achieved improved accuracy in automated disease classification using ensemble learning for medical image analysis. This trend extends to bioinformatics, where Nie et al.[20] explored the value of ensemble clustering methods for analyzing single-cell RNA sequencing data. These studies suggest that ensemble techniques offer a promising avenue for achieving better results in a range of fields.

Proposed methodology

The subsequent phases of the suggested methodology are shown in Fig. 1. After obtaining data from benchmark resource, it is trained for feature engineering, pre-processing, and resampling. Within the ensemble learning framework, we employ a stacking classifier after extracting the features for the proposed model. Our study focuses on distinguishing posts containing suicidal ideation from non-suicidal content. To accomplish this, we leveraged an ensemble learning approach, combining multiple machine learning classifiers to capture subtle distinctions in text that may indicate distress signals. The classification process comprises two stages:

Preprocessing and Feature Extraction: The text data underwent preprocessing to remove noise and enhance interpretability. Techniques included tokenization, stop-word removal, and word vectorization (using TF-IDF) to convert text into meaningful numerical features for training our model.

Ensemble Classification Approach: For the classification of suicidal and non-suicidal ideations, we employed a stacked ensemble model comprising four base classifiers—Support Vector Machine (SVM), Logistic Regression (LR), Gradient Boosting (GB), and Decision Tree (DT)—with a Random Forest model acting as the meta-classifier. This approach was selected to balance performance and interpretability, maximizing detection accuracy while ensuring that classifier behaviors remain transparent and explainable.

Rationale for classifier selection

Our classifier choices align with both the technical needs of suicide ideation detection and ethical guidelines for transparency and interpretability:

- **Support Vector Machine (SVM)**: Chosen for its effectiveness in handling high-dimensional data, SVM is known for its robustness in text classification, helping to differentiate nuanced language features[11].
- **Logistic Regression (LR)**: This linear model provides interpretable results and contributes to the ensemble by identifying straightforward linear relationships between features[12].
- **Gradient Boosting (GB)**: GB's iterative improvement capability helps to refine predictions based on errors in prior classifiers, capturing complex patterns associated with distressed language[13].
- **Decision Tree (DT)**: Known for its interpretability, DT allows us to trace classification paths, making it a critical component in explaining model decisions and building trust in the system[14].

The Random Forest meta-classifier[15] integrates predictions from these classifiers, leveraging diverse perspectives on the data to enhance predictive accuracy. Additionally, as each classifier within the ensemble offers varying interpretability levels, this approach ensures that we achieve both model performance and transparency, allowing stakeholders—such as mental health professionals—to understand and trust the decisions made.

Ethical Reporting Compliance

Adhering to ethical standards, this study includes:

- **Transparency in Classifier Functions**: By explaining the purpose and role of each classifier, we enhance the study's interpretability, aligning with explainable AI (XAI) principles.
- **Non-Specific Descriptions of Suicidal Language**: In line with reporting guidelines, sensitive information is summarized without direct quotations or identifiable content.

Through these adjustments, the model adheres to the principles of ethical AI research, especially in sensitive domains like mental health, where model interpretability and ethical reporting are critical.

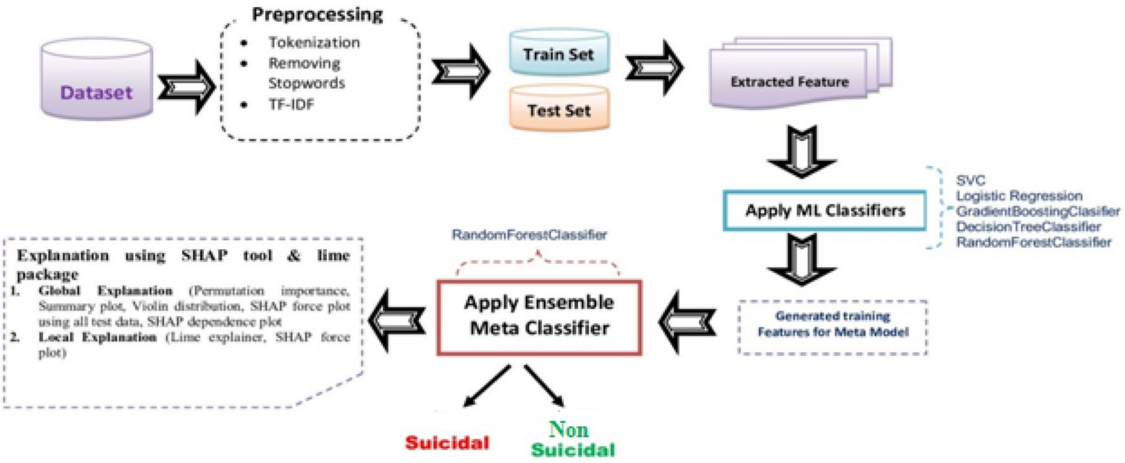

Dataset
Preprocessing
Tokenization
Removing Stopwords
TF-IDF
Train Set
Test Set
Extracted Feature
Apply ML Classifiers
SVC
Logistic Regression
GradientBoostingClasifier
DecisionTreeClassifier
RandomForestClassifier
Generated training Features for Meta Model
RandomForestClassifier
Apply Ensemble Meta Classifier
Explanation using SHAP tool & lime package
1. Global Explanation (Permutation importance, Summary plot, Violin distribution, SHAP force plot using all test data, SHAP dependence plot
2. Local Explanation (Lime explainer, SHAP force plot)
Suicidal
Non Suicidal

Fig. 1. Proposed system overview.

Dataset acquisition

The data being studied comes from an collection of Reddit posts that can be found on Kaggle[21]. It consists of a total of 12,638 posts split between those discussing thoughts (6610posts). Those focusing on other topics (6028 posts). In this context the term "SuicideWatch" go beyond its meaning related to monitoring in controlled settings like prisons, hospitals, mental health centers and military installations. On RedditSuicideWatchserves as a space where people who may be feeling distressed or having thoughts can find support from the community share their stories and express themselves anonymously[22].

This dataset not offers insights for studying language use and themes related to suicidal thoughts but also presents a mix of direct accounts of personal struggles and supportive or general content. The equal representation of both types of posts provides a rounded foundation, for machine learning models. Analyses seeking to detect linguistic patterns, emotional tone and potential signs of suicidal ideation. This dataset is incredibly valuable, for studying conversations about health crises due to its thoroughness and sensitivity. By examining this data we can uncover language patterns and themes that might be linked to thoughts of suicide. This insight could help create automated systems, for detecting and addressing these issues on[23]. Table 1 shows details of the acquired dataset.

Dataset overview and ethical considerations

The dataset used in this study consists of anonymized Reddit posts from the *SuicideWatch* and related subreddits, publically available via Kaggle. To ethically handle sensitive data on suicidal ideation, this study follows current recommendations on reporting such material responsibly[1]. In accordance with ethical standards, no individual post content is shown, and posts are instead summarized to avoid potential distress to readers. Labels were assigned based on detected patterns in language use, not as definitive assessments of intent, as part of our commitment to maintaining objectivity and sensitivity in studying mental health data.

Rationale for labelling the Analysed Posts

The rationale for labeling posts as suicidal or non-suicidal, it is important to clarify that the dataset used in this study—a benchmark, publicly available dataset from Kaggle—comes pre-labeled with these classifications[21]. This dataset includes posts from Reddit's Suicide Watch and related subreddits, curated to support research on mental health by distinguishing posts that express suicidal ideation from those that do not. Our analysis leverages these pre-assigned labels, which are based on linguistic patterns and themes identified in prior studies as indicative of distress or ideation[22,23]. Following ethical guidelines for sensitive content as outlined in ReportingOnSuicide.org, our study respects the complexities involved in handling mental health data by maintaining objectivity and avoiding any sensationalism. The labels serve solely as a tool for categorizing language use patterns without implying definitive intent, ensuring sensitivity and ethical responsibility in our approach. This allows us to develop automated systems that can recognize indicators of distress while adhering to best practices in mental health research and reporting.

Pre-processing and data preparation

Following pre-processing steps are applied on the acquired dataset: (1) Text processing, (2) Re-sampling, (3) Training and Testing.

1) Text Processing We employed different text processing techniques, such as tokenization, and stop-word removal.

Tokenization: The Python NLTK tokenizer is used in this approach to segment terms inside the text.

Removing Stop Words: Python's NLTK is used to remove stop words, which are common terms like "a," "the," "or," and so on. This is carried out to keep every word in the dataset as meaningful as feasible.

2) Re-Sampling approach When a model is trained on a skewed or imbalanced distributed dataset, the issue of class imbalance frequently occurs, and predictions for smaller class scales are frequently ignored. As a result, we attempted to stabilize this data sparsity. To balance the classes, we used a Re-sampling technique. This decreased inconsistency among the classes of each group and effectively handled the model's performance[24].

Data level re-sampling approach: To balance the instances of all classes, data preparation strategies that focus on rescaling the training datasets must be used. Oversampling and under-sampling are the two main methods for class resizing. Oversampling is the process of generating fresh samples for minority classes in order to attain a more equitable distribution across class samples. The random oversampling strategy includes

Name	Suicidal ideation-Reddit dataset
Description	Dataset is acquired from Kaggle by using The dataset comprised of 12638posts
Instances	12,638
Suicidal posts	6610 (52.303%)
Non-suicidal Posts	6028 (47.697%)
Format	Text
Default task	Suicidalor non-suicidal ideations classification from posts
Updated	2019
Origin	Kaggle
Size	04 MB
Owner	Varun

Table 1. Details of dataset.

reproducing random minority class samples in the training dataset and increasing their numbers to match the majority class level[24].

Random oversampling: The process of randomly duplicating samples from the minority class is repeated until the necessary balance is achieved, effectively increasing the number of samples in the minority class.

***3) Training and testing data*:** The primary datasets required are the training set, which is used to develop the model, and the testing set, which is used to evaluate the system's performance. For this purpose, we just used the *train_test_split* function[25].

Training dataset: The training dataset is mostly used in the proposed model's training. We used 70% of the available dataset for training to allow for variability. The training phase of the model is made up of data with labels that are already known to the classifier, and this labelled data is used to develop the model[26]. [1]https://reportingonsuicide.org/wp-content/uploads/2024/04/ROS-One-Pager-updated-2024.pdf

Test dataset: The testing dataset includes samples that were not included in the training dataset, with 30% of the data selected for conclusive model evaluation. This unnamed test dataset is used for model evaluation after training. Predictions are created for each occurrence in the test dataset using ensemble techniques, and the model's accuracy and efficiency are measured by comparing the anticipated values to the actual values[24,27].

Feature engineering

Feature engineering includes following modules.

1) CountVectorizer: Tokenization is a procedure used to turn text or documents into numerical matrices. The count vector is a popular encoding method that is implemented by software such as CountVectorizer, which converts documents into word vectors. CountVectorizer, commonly known as a document-term matrix or Bag of Words, generates a matrix containing document and token counts. Each sentence or document is divided into tokens, and the count of each token occurring in a message is added. CountVectorizer can tokenize an entire text document, create a dictionary of defined terms, and encode new documents using a pre-defined vocabulary[20,28].

2) Term Frequency: Term Frequency (TF) reflects the weight of a word or phrase based on how frequently it appears in a document.

$$Tf\left(t,d\right)=\frac{countoftermind}{number\ of\ \ words\,in\,d}, \tag{1}$$

3) Inverse document Frequency: Inverse Document Frequency (IDF) is a numerical representation that measures the rarity or uniqueness of a term within a collection of documents. It is calculated by taking the logarithm of the inverse of the fraction of documents containing the term. In essence, IDF highlights the importance of a term by assigning higher weights to terms that are less common across the entire document set. This weighting scheme is commonly used in text processing and information retrieval to emphasize the significance of terms that are more discriminative or rare[6]. It is computed as follows:

$$IDF\left(t,D\right)=log\left(\frac{N}{DF+1}\right), \tag{2}$$

DF(t,D) is the Document Frequency; it indicates the number of documents in D that contain the term *t. t* is the term or word; *D* is the collection of documents; N is the total number of documents in D.

4) Term Frequency Inverse Document Frequency: The Term Frequency-Inverse Document Frequency *(TF-IDF)* statistic measures the importance of a term in a document in comparison to its total occurrence in a collection of documents. It joined two measures: Term Frequency (TF) and Inverse Document Frequency *(IDF), TF* determines how often a term appears in a document, and IDF evaluates the uniqueness or rarity of a term in the entire document set. The *TF-IDF*value is computed by multiplying the values of TF and IDF, these results in a weighted representation of phrases that are common in documents and differ on the corpus.

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

where t is the phrase or word, The document is denoted by d, and D is a document collection. The Term Frequency *(TF(t,d)* represents the number of times the term t appears in the document.

Ensemble learning mathematical formulation for suicidal ideations recognition

Ensemble learning can be used in the context of suicidalor non-suicidal Ideations classification on the benchmark dataset to integrate the predictions of suicidal or non-suicidal Ideations classification models to achieve a more accurate prediction of a person's suicidal or non-suicidal Ideations (see Fig. 2).

The general formula for ensemble learning is as follows:

$$E=f(M1(X),M2(X),...,Mn(X)... \tag{4}$$

where E = Predicted_Suicidal_Ideation; F = The ensemble function, *f*, aggregates the predictions of the individual models; and *M1(X), M2(X),…, Mn(X)* are the predictions of the individual models for the input data point X.

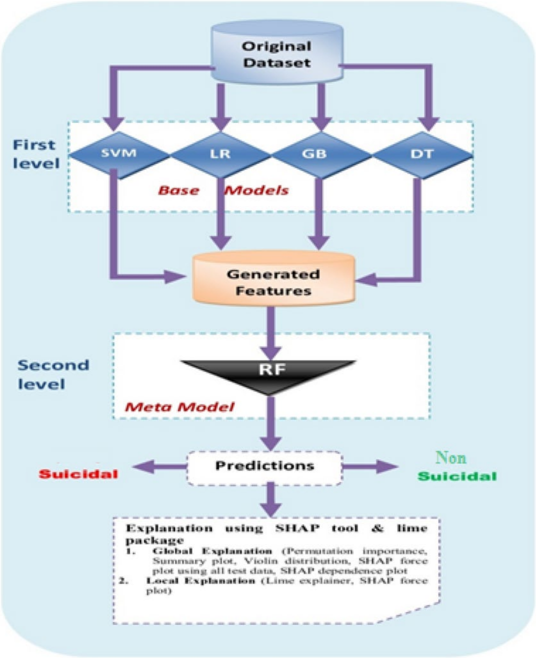

Original Dataset
First level
SVM
LR
GB
DT
Base Models
Generated Features
Second level
RF
Meta Model
Suicidal
Predictions
Non Suicidal
Explanation using SHAP tool & lime package
1. Global Explanation (Permutation importance, Summary plot, Violin distribution, SHAP force plot using all test data, SHAP dependence plot
2. Local Explanation (Lime explainer, SHAP force plot)

Fig. 2. Proposed 2-layer Stacking Ensemble Method.

In our case, consider a series of base models (SVM, LR, GB, DT), each of which has been trained to predict the Suicidalor Non Suicidal Ideations type using a set of person's sentiments expressed in social media posts. Let's call the set of base models $M = M1, M2, \ldots, Mn$, where Mi is the *ith* base model.

Each base model's predictions can be blended using a variety of ensemble approaches, each with its own mathematical formulation. In this work, weemploy stacking approach. Stacking is a robust ensemble learning technique in which a meta-model is trained to aggregate the predictions of numerous base models in order to improve overall classification performance. Stacking can efficiently capture the intricacies of suicidal ideations and improve the accuracy of prediction in the context of suicidal ideation classification on the benchmark dataset[22].

The stacking ensemble method for Suicidalor Non-Suicidal Ideations classification involves integrating predictions from different base models. The formulation can be stated mathematically as follows:

- The training dataset is represented by $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ as the training dataset, where x_i is the i^{th} example's feature vector and $\mathrm{y_i}$ is its matching Suicidal Ideationstype label (suicidalor non-suicidal).
- M, where $M = 4$, is the number of base classifiers. The base classifiers are denoted by the letters f_1, f_2, f_3, f_4 , which stand for SVM, decision tree, logistic regression, and gradient boosting, respectively.
- g as the meta-classifier, which in this case is a random forest

1. Training Phase of Base Classifiers:

For every base classifier $m = 1,2,3,4$:

$$f_m \leftarrow train\,(D)\,,$$

For each case i, obtain predictions from each basic classifier:

$$h_{m,i} = f_m\,(x_i)\,form = 1,2,3,4,$$

2. Development of Meta-features:

Create a new dataset 'D' with each sample i^{th} including the original features $x_{i,}$ and the predictions from each basic classifier:

$$(x_i, h_{1,i}, h_{2,i}, h_{3,i}, h_{4,i}, y_i)\,,$$

3. Random Forest-Based Meta-Classifier Training:

$$g \leftarrow train\,\left(D'\right),$$

The meta-classifier (random forest) is trained to learn a mapping from the base classifier predictions to the true labels.

4. Prediction Phase:

Obtain predictions from each base classifier given new input x:

$$h_m\left(x\right) = f_m\left(x\right) form = 1.2.3.4,$$

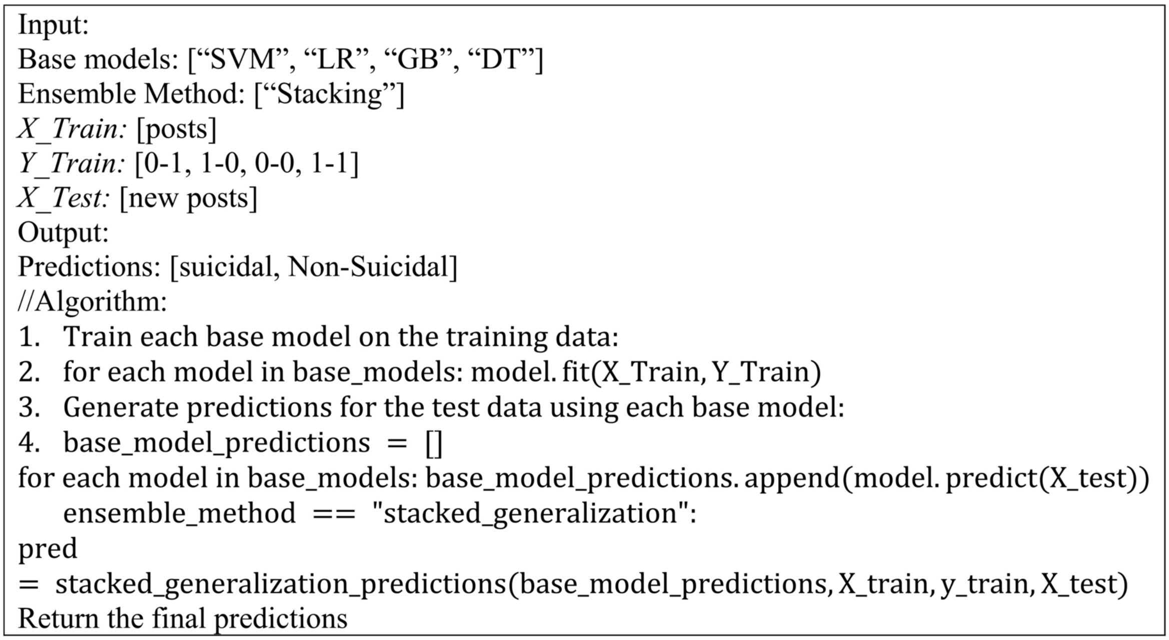
To acquire the final prediction, create a new instance $\left(x, h_1\left(x\right), h_2\left(x\right), h_3\left(x\right), h_4\left(x\right)\right)$ and utilise the learned meta-classifier (random forest).

$$g\left(x\right) = predict\left(g, \left(x, h_1\left(x\right), h_2\left(x\right), h_3\left(x\right), h_4\left(x\right)\right)\right).$$

To summarize, the mathematical formulation entails training each base classifier, generating predictions, creating a new dataset with meta-features, training the meta-classifier (random forest), and making predictions using the ensemble.

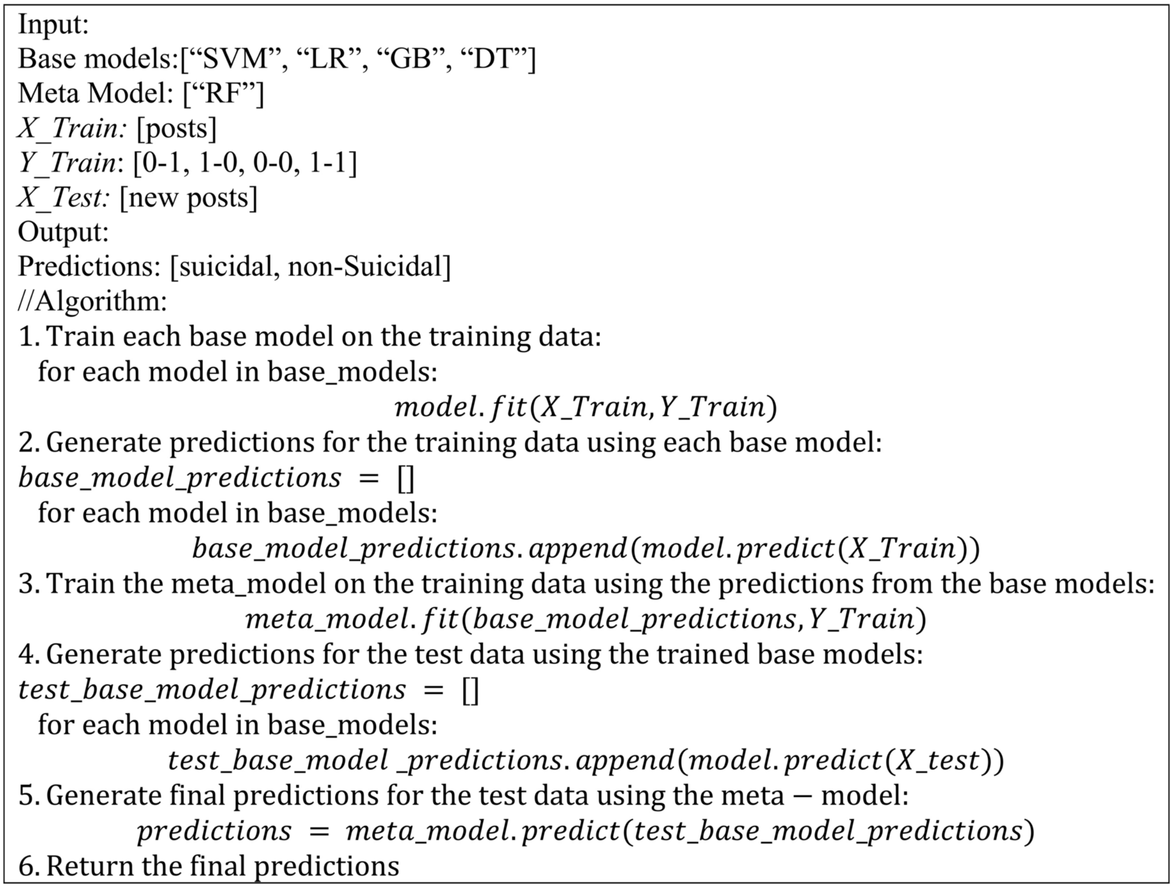
The stacking model mixes numerous machine learning models to improve forecasting accuracy. The choice of base-models and the Meta-model during development has a substantial impact on the performance of the stacking ensemble classifier. The basic assumption in adopting a stacking model is that the base-models have good prediction performance and can be used to improve classification results. SVM, Gradient Boosting, Logistic regression as well as Decision Tree were the base models in this research. As a general machine learning model, SVM is capable of performing linear and nonlinear classification, and is particularly suitable for complex data and small and medium-sized data sets. Logistic regression as a classification model is useful for identifying project outcomes because it is easy to implement, analyze and train. Gradient boosting is equivalent to gradient descent technology and is a powerful machine learning algorithm. It is a comprehensive and powerful approach to improvement because it excels at discovering the best solutions in every situation. Decision trees are a well-known machine learning technique that uses a tree structure to describe data to solve machine learning issues. It is highly adaptable as it can be used to solve both classification and regression problems. The predictions of various classifiers selected as base classifiers are used to train a meta-model that works for new features. Finally, an ensemble learning strategy based on stacking is used. SVM, LR, GB, and DT classifiers are used as base models in this study, while Random Forest is used as the Meta model in the second level.

Algorithms 1 and 2 shows overall system's working and stacking classifier working respectively.



```
Input:
Base models: ["SVM", "LR", "GB", "DT"]
Ensemble Method: ["Stacking"]
X_Train: [posts]
Y_Train: [0-1, 1-0, 0-0, 1-1]
X_Test: [new posts]
Output:
Predictions: [suicidal, Non-Suicidal]
//Algorithm:
1.  Train each base model on the training data:
2.  for each model in base_models: model. fit(X_Train, Y_Train)
3.  Generate predictions for the test data using each base model:
4.  base_model_predictions = []
for each model in base_models: base_model_predictions. append(model. predict(X_test))
    ensemble_method == "stacked_generalization":
pred
= stacked_generalization_predictions(base_model_predictions, X_train, y_train, X_test)
Return the final predictions
```

Algorithm 1 Overall working of the proposed System.



```
Input:
Base models:["SVM", "LR", "GB", "DT"]
Meta Model: ["RF"]
X_Train: [posts]
Y_Train: [0-1, 1-0, 0-0, 1-1]
X_Test: [new posts]
Output:
Predictions: [suicidal, non-Suicidal]
//Algorithm:
1. Train each base model on the training data:
  for each model in base_models:
                    $model.fit(X\_Train, Y\_Train)$
2. Generate predictions for the training data using each base model:
$base\_model\_predictions = []$
  for each model in base_models:
          $base\_model\_predictions.append(model.predict(X\_Train))$
3. Train the meta_model on the training data using the predictions from the base models:
             $meta\_model.fit(base\_model\_predictions, Y\_Train)$
4. Generate predictions for the test data using the trained base models:
$test\_base\_model\_predictions = []$
  for each model in base_models:
          $test\_base\_model\ \_predictions.append(model.predict(X\_test))$
5. Generate final predictions for the test data using the meta − model:
          $predictions = meta\_model.predict(test\_base\_model\_predictions)$
6. Return the final predictions
```

Algorithm 2 Stacking Working Process.

Explanations of ML models using explainable AI (XAI) approach

Explainable AI (XAI) in detecting thoughts especially when using methods, like Random Forests or Gradient Boosting is a merging point between machine learning and mental health support. The model checks amounts of information, like text from the media or clinical records, to discover patterns suggestive of suicidal thoughts or behaviors. However, the complexity of models, which blend algorithmic predictions to improve accuracy, often presents challenges known as "black box" problems. It shows that activity of decision-making has become difficult and vague to understand[4].

Explainable Artificial Intelligence (XAI) objective to address this obstacle v.i.a transparency improving and clarity of this model. Its purpose is to gain insight into the reasoning behind prediction of model. Attribute prediction terminology that using input features, such as Local Interpretable Model Explanations (LIME) or Shapley Additive Explanations (SHAP). Such model give in-dept information about how much part of the data affects the model decision-making process. For Example, in text data certain words are phrases may help identify innovative idea and highlights these important elements with the help of XAI[4]. There are many reasons why using XAI for idea detection is important. To build human confidence in AI systems, it is important to ensure transparency in human decision-making so that model errors or biases can be identified. Additionally it plays a role in assisting health experts by offering interpretable evidence to inform interventions. Combining the abilities of models with the clarity offered by XAI allows researchers and clinicians to gain deeper insights into individuals, at risk thereby ensuring that interventions are both timely and tailored for maximum effectiveness[29].

Model Explanation using SHAP: Developing a framework, for SHAP (SHapley Additive exPlanations) within the realm of Explainable AI for identifying suicidal thoughts particularly in scenarios utilizing ensemble methods like a Random Forest involves determining the calculation of SHAP values tailored to this specific use case. The objective is to break down a models prediction into contributions from features offering an understanding of why the model arrived at a prediction regarding suicidal ideation based on input features such, as text data extracted from social media posts. Lets outline this procedure;

Ensemble Model Prediction: Let's say we have a model M (e.g. Random Forest) that receives an input vector x with each, xi denoting a feature extracted from text data (like $TF - IDF$ scores, etc.). Generate a prediction *p*. This prediction represents the likelihood of the input indicating thoughts of suicide.

SHAP Value Calculation: Below is an overview of how SHAP can be applied in Explainable AI (XAI) for identifying suicidal thoughts using an ensemble approach even though the detailed mathematical aspects might be complex.

The SHAP value, for a feature xi in a prediction is determined by assessing how much xi contributes to the variation between the models prediction for input x and the average prediction of the model, for all inputs. In terms when looking at a feature *xi* its SHAP value *ϕi* is calculated as follows;

$$\emptyset_i = \sum_{S\subseteq F\backslash\{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f\left(S\cup\{i\}\right)-f\left(S\right)\right]. \tag{5}$$

where:

Feature symbol	Description
F	set of all features
S	subset of features excluding i
$\mid S \mid$	cardinality of subset S, i.e., the number of features in S The size of subset S, which refers to the count of features, in S
$\mid F \mid$	total number of features
$f\left(S\right)$	prediction of the model when only the features in set S are included The models forecast is based on the elements, within S
$f\left(S\cup\{i\}\right)$	prediction of the model when the features in S plus the feature i are included The models forecast changes when incorporating the features, in set S along, with feature i
ϕi	SHAP value for feature i, representing its marginal contribution to the prediction The SHAP value, for feature "i" indicates its impact, on the prediction

Model Interpretation The prediction p of the model, for an instance can be broken down as;

$$p = \emptyset_o + \sum_{i=1}^{n} \emptyset_i$$

In this equation $\emptyset_o$ denotes the starting point, which reflects the models prediction, across the dataset. The sum of SHAP values for all features $\sum_{i=1}^{n} \varnothing_i$ accounts for how each feature influences the deviation, from the starting point value.

Application to Suicidal Ideation Detection When suicidal detecting thoughts model M assigns a higher likelihood p to cases, with increased risk. The SHAP values $\varnothing_i$ for each feature xi, such as words, phrases or sentiment scores show how influence each feature has on the models final prediction. A positive $\varnothing_i$ indicates a feature that boosts the models prediction of thoughts while a negative $\varnothing_i$ suggests a feature that reduces it. This breakdown assists health professionals in grasping the AI models reasoning potentially highlighting factors to consider in assessment and intervention planning. This mathematical approach lays the groundwork, for implementing SHAP in Explainable AI for identifying thoughts providing clarity and insights into the mechanisms of sophisticated ensemble methods.

Applied example

Let us assume a simplified scenario with a small dataset and focus on suicidal or Non-Suicidal Ideations classification (e.g., suicidal vs. non-suicidal) using the benchmark dataset for simplicity. SVM, decision tree, logistic regression, and gradient boosting are the base classifiers.

1) ***Preparation of Data:*** Consider a dataset containing a few individuals, each labelled with their Suicidal Ideations and associated attributes.
2) ***Data Segmentation:*** Divide the dataset into two parts: training and testing (Fig. 3).
3) ***Fundamental Classifiers:*** Make predictions on the test set after training each basic classifier (SVM, decision tree, logistic regression, gradient boosting) on the training set.
4) ***Developing Meta-features:*** To build meta-features for the stacking classifier, gather the predictions from every base classifier.
5) ***Stacking Classifier:*** Develop a stacking classifier by labeling the meta-features with suicidal ideation and classes.

Ensemble-based Prediction: Use the stacking classifier to make predictions on new data by combining the predictions of the base classifiers.

This simplified example shows the essential stages involved in ensemble-based Suicidalor Non-Suicidal Ideations from Social Media Text on the benchmark dataset using the stacking classifier with four base classifiers.

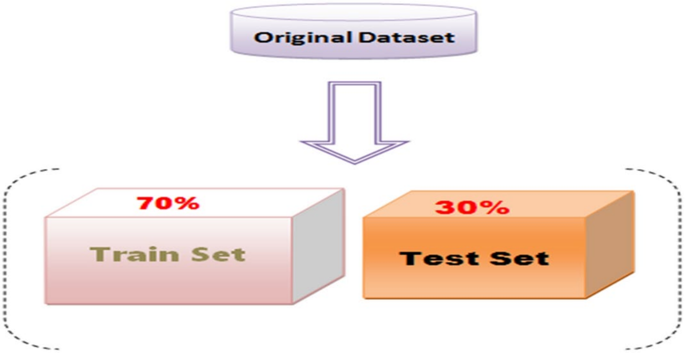
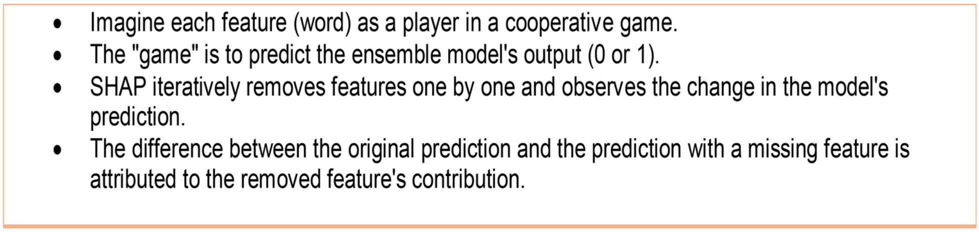


Fig. 3. Dataset partitioning.



- Imagine each feature (word) as a player in a cooperative game.
- The "game" is to predict the ensemble model's output (0 or 1).
- SHAP iteratively removes features one by one and observes the change in the model's prediction.
- The difference between the original prediction and the prediction with a missing feature is attributed to the removed feature's contribution.

Fig. 4. SHAP values calculation.

6) ***A Comprehensive Mathematical Exposition of XAI (SHAP), to Detect Suicide Thoughts:*** In a situation where we have a model that categorizes social media posts as either expressing suicidal thoughts (1) or not (0) we aim to employ SHAP to clarify why the model predicted a particular post.
7) **Preprocessing** (Example, for Illustration); Different preprocessing procedures are applied as in the instance (tokenization converting to lowercase removing stop words).

Representation of feature vectors; In this scenario each word is given an index according to the vocabulary turning the post into a feature vector;

Calculating SHAP Values; SHAP employs a game theory strategy to allocate the prediction of the combined model, across all features (words) in the post. Here's a simplified example (See Fig. 4);

Results and discussions

A set of results generated from the proposed model, are declared in this section by thoroughly answering the research questions.

Answer to RQ.1

To answer Research Question 1 (RQ1), ensemble learning is used to predict suicidal ideation from the input text. The stacking classifier was chosen as the ensemble learning method because of its efficacy as a well-organized heterogeneous ensemble methodology. Stacking works as a super multi-layer model, with each layer consisting of one or more models, and the outcomes of the second layer influenced by the models of the previous layer. The first-level base models, which include SVM, Decision Tree, Logistic Regression, and Gradient Boosting, generate predictions that are used to train the Meta classifier in the second level, which acts as a new feature. Random Forest, the final Metamodel, refines the forecasts. Using default parameters and experimenting with other parameter combinations in both the Base models and the Metamodel yielded a variety of results.

1. **Parameter Setup** Table 2 shows the setting of the Parameter for the different classifiers
2. **Algorithmic Complexity of the Proposed System** The complexity of ensemble-based suicidal ideation classification on the suicidal Ideation dataset using the stacking classifier can be computed as follows, assuming '*n*' data points and 4 base models:

Training base models: Depending on the particular model being used, different base models have different training costs. The complexity of SVM and Decision Trees is usually $O(n^2)$, where n is the number of data points. The complexity for logistic regression and gradient boosting is usually $O(n * d)$, Where d is the number of features.

In this instance, let's assume that each base model's training has an average complexity of $O(n^2)$. consequently, the training of all four base models has a total complexity of:

Model	Parameters	Description
SVM	*Kernel* = *linear*	For making training dataset linearly detachable It converts it into higher dimensions
	C (Regularization) = 100,50,0.025	This parameter describes that how much jumbling of training data is acceptable in the model
	Gamma = *0.2, auto*	This parameter decides how much influence the data points at a certain distance from the hyper plane will have
	Random_state = 42	It manages the dragging process
DT	*max_depth* = 2, 6	Best possible depth of the tree is decided by this method
	Random_state = 42	It manages the dragging process
LR	*C* = 100, 200, 2000	This parameter has numbers that instruct the model what to do with the features. Same as in svm smaller values define powerful regularization
GB	*min_samples_leaf* = 2	It determines the littlest samples called observations. This needed in a final node or leaf. It can be utilized for over fitting
	max_depth = 5	This parameter is tuned for excellent performance The highest depth limits the number of nodes in the tree
	n_estimators = 500	It describes the number of trees in the forest. Generally the bigger number of trees can learn the data more excellently
RF	*max_depth* = 10	It represents the splits so that every decision tree number of splits that each decision tree is permitted to build
	max_features = auto	It represents the column numbers that are displayed to every decision tree
	min_samples_split = 0.005	The least samples needed to split an interior node
	min_samples_leaf = 0.005	The least possible samples needed to be at a leaf node
	n_estimators = 10	It describes the number of trees in the forest. Generally the bigger number of trees can learn the data more excellently
	n_jobs = − 1	It indicated the jobs to run in comparable in fit and predict
	random_state = 42	Controls both the randomness of the bootstrapping of the samples used when building trees

Table 2. Parameter setting of applied classifiers.

1. SVM: gamma = scale, C = 1.0, kernel = rbf
2. Logistic Regression = C = 1.0, penalty = l2, solver = lbfgs
3. Gradient Boosting: loss = deviance, learning_rate = 0.1, n_estimators = 100
4. Decision Tree: criterion = gini, splitter = best, max_depth = None, min_samples_split = 2

Table 3. Parameters for weak learners.

		Weak Classifiers					
Metrics		SVM	LR	GB	DT	Proposed Stacking Classifier	Average Score of Proposed Model
Accuracy	Suicidal	0.95	0.88	0.85	0.90	0.95	92%
	Non-Suicidal	0.98	0.93	0.87	0.94	0.99	
Precision	Suicidal	0.94	0.88	0.85	0.86	0.97	92%
	Non-Suicidal	0.99	0.95	0.88	0.99	0.99	
Recall	Suicidal	0.95	0.88	0.85	0.96	0.94	91%
	Non-Suicidal	0.96	0.92	0.88	0.88	0.99	
F1-Score	Suicidal	0.95	0.89	0.85	0.91	0.955	91%
	Non-Suicidal	0.98	0.94	0.88	0.94	0.99	

Table 4. Results of weak classifiers and proposed model with default parameter settings.

$$4 * O(n^2) = O(4n^2)$$

Training stacking classifier: Using the base models' predictions as input, the stacking classifier learns how to combine them to produce predictions that are more accurate. Depending on the particular model being utilized, the stacking classifier's training difficulty varies. $O(m * d)$, where m is the number of base models and d is the number of features, is the usual complexity for logistic regression.

Let's assume that the stacking classifier's training complexity in this instance is $O(m * d)$. Consequently, using four base models to train the stacking classifier has the following complexity: $O(4 * d) = O(4d)$

$O(p)$ is the complexity of XAI module.

The total level of complexity: The total complexity of ensemble-based suicidal ideation categorization on the suicidal dataset with '*n*' data points and four base models is: *O(4n^2)* + *O(4d)* + *O(p)*. Because *n* is often substantially larger than *d*, *O(4n^2)* is the major term in the complexity. As a result, the ensemble method's overall complexity can be estimated as *O(4n^2)* + *O(p)*.

Parameters	Metrics		SVM	LR	GB	DT	Stacking	Average Score
SVM *random_state* = 42 LR: *random_state* = 42 GB: *random_state* = 42 *n_estimators* = 1000 DT: *random_state* = 42 Stacking(*cv* = 3, *estimators* = *[('SVM,LR,GB,DT* *final_estimator* = *RF(random_state)* = 42, *Max_depth* = 10, *Min_samples_leaf* = 0.005 *Min_samples_split* = 0.005 *Max_features* = auto *n_jobs* = -1, *random_state* = 42	Accuracy	Suicidal	0.95	0.87	0.94	0.86	0.96	92%
		Non-Suicidal	0.85	0.82	0.85	0.77	0.85	
	Precision	Suicidal	0.94	0.88	0.91	0.84	0.93	90%
		Non-Suicidal	0.99	0.94	0.95	0.98	0.97	
	Recall	Suicidal	0.94	0.87	0.96	0.95	0.94	92.00%
		Non-Suicidal	0.95	0.90	0.92	0.85	0.94	
	F1-Score	Suicidal	0.93	0.87	0.92	0.88	0.94	91%
		Non-Suicidal	0.97	0.91	0.95	0.94	0.97	

Table 5. Different results while using different parameters.

Parameters	Metrics		SVM	LR	GB	DT	Stacking	Average Score
SVM: *C* = *50* *Gamma* = *auto* *Probability* = *true* *Degree* = *1* *LR:* *C* = *200* *GB:* *Max_features* = *0.2* *Max_depth* = *10* *Min_samples_leaf* = *2* *DT:* *Max_depth* = *6* *random_state* = *42* *StackingClassifier* *cv* = *3, estimators* = *[('SVM,LR,GB,DT* *final_estimator* = *RF(Max_features* = *auto* *Max_depth* = *10* *Min_samples_leaf* = *0.005* *Min_samples_split* = *0.005* *n_jobs* = *-1* *n_estimators* = *10* *random_state* = *40*	Accuracy	Suicidal	0.52	0.90	0.91	0.73	0.94	92%
		Non-Suicidal	0.70	0.94	0.95	0.81	0.94	
	Precision	Suicidal	0.92	0.87	0.91	0.76	0.95	91%
		Non-Suicidal	0.97	0.98	0.98	0.79	0.98	
	Recall	Suicidal	0.73	0.94	0.92	0.75	0.96	94%
		Non-Suicidal	0.48	0.93	0.93	0.83	0.97	
	F1-Score	Suicidal	0.67	0.90	0.91	0.72	0.94	91%
		Non-Suicidal	0.62	0.95	0.95	0.81	0.91	

Table 6. Different results while using different parameters.

Answer of RQ.2

Answer toward RQ.2: what is the efficiency of ensemble methods as compared to weak learners (classifiers)?

We have generated results by using default parameter settings for weak classifiers and then compared them with our proposed stacking classifier. Here are the default parameters for weak classifiers (Table 3), whereas Table 4 shows results of different classifiers.

The results of many weak machine learning classifiers, as well as our suggested stacking classifier, are presented here. Our stacking classifier's findings were acquired by using weak classifiers as the basis learners. Notably, using a re-sampling strategy on our benchmark dataset, we achieved outstanding results across all classifiers without using special parameter values. When compared to the weak classifiers, our suggested classifier performed better across all suicidal ideations. When compared to the weak classifiers, the stacking classifier attained an amazing 99% accuracy in N-S characteristics. To improve results, we will investigate parameter settings for both the base models and the meta-model.

We used a variety of parameter setups, which resulted in higher recall levels when compared to the results obtained with the default parameter settings. All base classifiers had a consistent random state of 42, while the Gradient Boosting (GB) classifier had*n_estimators* set to 1000. We investigated several parameter values in the Random Forest (RF) meta-model. We also tweaked the Stacking classifier with cv = 3, yielding a high recall accuracy of 92.00% (see Table 5).

We introduced various parameter values for both the base classifiers and the meta-model in Table 6. This change resulted in an enhanced average recall accuracy of 93.50%, which outperformed the previous findings. However, the average precision accuracy has reduced significantly to 90.30%, compared to the prior precise average accuracy of 91%.

Table 7 shows results obtained using another version of parameters applied in different classifiers.

We provided additional parameter changes for both the basis classifiers and the meta-model in Table 7 above. However, in this case, we just applied the parameter 'c' to SVM and LR, resulting in less favorable results than the previous table findings. The decrease in recall scores for the SVM classifier in Tables 6 and 7 can be attributed to

Parameters	Metrics		SVM	LR	GB	DT	Stacking	Average Score
SVM: *C = 0.025* *LR:* *C = 2000* *GB:* *N_estimators = 500* *Max_depth = 5* *Min_samples_leaf = 2* *DT:* *Max_depth = 6* *random_state = 42* *StackingClassifier* *cv = 3, estimators = [('SVM,LR,GB,DT* *final_estimator = RF(* *Max_depth = 6* *Min_samples_leaf = 2* *n_estimators = 500* *n_jobs = -1* *warm_start = true*	Accuracy	Suicidal	0.51	0.90	0.92	0.77	0.94	91%
		Non-Suicidal	0.61	0.95	0.94	0.79	0.95	
		Suicidal	0.52	0.83	0.84	0.77	0.90	
		Non-Suicidal	0.56	0.76	0.84	0.76	0.90	
		Suicidal	0.34	0.91	0.90	0.95	0.97	
		Non-Suicidal	0.97	0.84	0.84	0.71	0.76	
	F1-Score	Suicidal	0.41	0.92	0.93	0.77	0.95	90%
		Non-Suicidal	0.65	0.95	0.96	0.81	0.97	

Table 7. Different results while using different parameters.

Metric	Description	Ensemble technique (Random Forest and LSTM)	Ensemble technique (XGBoost& CNN)
Accuracy	Overall proportion of correctly classified samples	0.872	0.891
Precision (suicidal class)	Proportion of correctly identified suicidal posts among predicted suicidal posts	0.825	0.857
Recall (suicidal class)	Proportion of correctly identified suicidal posts out of all actual suicidal posts	0.783	0.814
F1-score (suicidal class)	Harmonic mean between Precision and Recall (Suicidal Class)	0.803	0.835
AUC-ROC	Area Under the ROC Curve	0.9	0.92
Specificity	Proportion of correctly identified non-suicidal posts among predicted non-suicidal posts	0.919	0.925
Confusion matrix	Breakdown of true positives, false positives, true negatives, and false negatives		
Time to explain prediction (avg.)	Average time taken by the model to explain a specific prediction	2.1 s	1.7 s

Table 8. Experimental results obtained on dataset#2.

SVM performance's sensitivity to the hyperparameter 'C'. The 'C' value also effects the bias-variance trade-off, which can lead to discrepancies.

Additional Experiments on Dataset#2:We conducted additional experiments on an other publically available dataset[30] dataset is made up of entries, from the “SuicideWatch” and “depression” sections on Reddit gathered using the Pushshift API. Entries from “SuicideWatch” range from December 16 2008 to January 2 2021 and are categorized under suicide. Entries from “depression” span from January 1 2009, to January 2 2021, are classified as depression. Additionally, posts not related to suicide were collected from the r/teenagers subreddit. The dataset is comprised of 232,074 unique values (see Table 8).

The results table suggests both ensemble techniques perform well in classifying suicidal and non-suicidal ideations. Here's a breakdown: (i) Overall Accuracy: Both models achieve high accuracy (above 87%), indicating a good ability to distinguish between the classes. However, a slight difference suggests Ensemble Technique (XGBoost & CNN) might be marginally better for general classification, (ii) Suicidal Class Performance: There's a trade-off between precision and recall for the suicidal class. Ensemble Technique (Random Forest & LSTM) has higher precision (fewer false positives), but Ensemble Technique (XGBoost & CNN) might miss fewer actual suicidal posts (higher recall). This highlights the importance of considering both metrics. The F1-score, which balances them, slightly favors Ensemble Technique (XGBoost & CNN), (iii) AUC-ROC: Both models show good ability to differentiate classes, with Ensemble Technique (XGBoost & CNN) performing marginally better (closer to 1), and (iv) Explainability: Ensemble Technique (XGBoost & CNN) seems to provide explanations faster, which is a benefit for interpretability. Overall, the results are promising, with both techniques demonstrating potential for real-world application. However, further analysis is needed:

Answer of RQ.3

RQ.2 “What is the performance of the proposed ensemble methods for suicidal ideation prediction with regard to baseline studies?”.

We evaluated our proposed model, against two methods referred to as[3,4]. The study by[3] focused on classifying suicidal thoughts using a recognized dataset (refer to Table 9). They employed learning in their methodology utilizing gradient boosting decision trees and SVM. Reported accuracy rates for their model were 68.3% for suicidal cases and 85% for simulated cases of non-suicidal ideation. Additionally, they outlined plans, for research aiming to develop a network that captures various forms of interactions.

Study	Aim	Technique	Dataset	Results (F1-core)

Liu et al.[3]	Suicidal ideation classification	Ensemble learning	Suicidal benchmark dataset	Suicidal = 70%
				Non-Suicidal = 84%

Malhotra et al.[4]	Suicidal ideation classification	XAI-based approach with Transformer	Suicidal benchmark dataset	Suicidal = 86%
				Non-Suicidal = 90%

Proposed (our work)	Suicidal ideation classification	XAI-based approach with Ensemble Stacking Classifier
		Base Models (SVM,LR,GB,DT)
		Meta Model(RF)
			Suicidal benchmark dataset	Suicidal = 95.5%
				Non-Suicidal = 99%

Table 9. Different results while using different parameters.

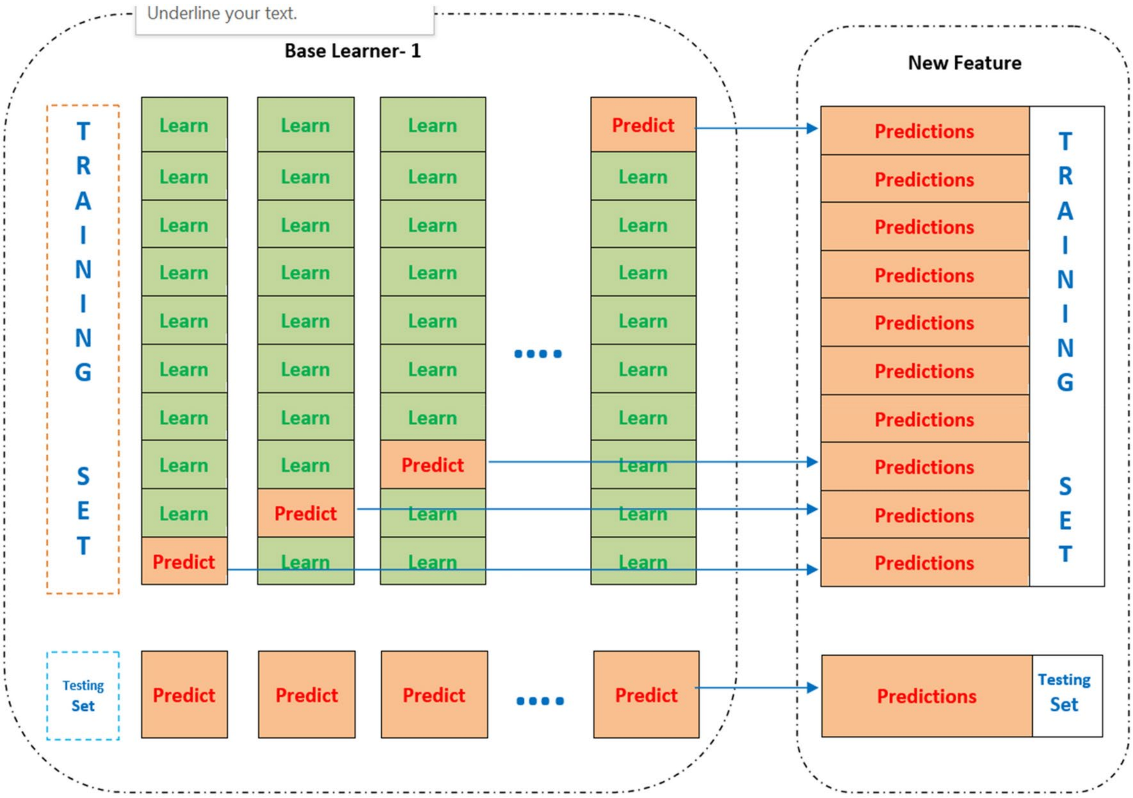


Fig. 5. Cross-validation model utilized in this investigation as a Base-learner.

Researchers[4] utilized a transformer based method along, with a dataset to effectively identify user's thoughts of suicide from the provided text. They attained recall rates of 88% for actual suicidal cases and 80%, for non suicidal instances resulting in an overall recall of 87%. The reported accuracy and f1 score both stood at 87%. Nevertheless when considering all aspects the comprehensive accuracy was determined to be 85%.

Proposed Work (our): In our proposed study, we used XAI-based Ensemble driven approach to predict suicidal ideation using the benchmark dataset. We started with re-sampling and then used super learning techniques within the ensemble, including the Stacking Classifier. The results for suicidal ideation were notably positive, demonstrating significant improvement over earlier results. Given the irregular nature of the dataset, we used resampling strategy, which produced significantly better results than any previous findings.

In Table 9, we compared our findings to the baseline investigations. Baseline research classified suicidal ideations using the same information. In comparison to baseline testing, our results are more accurate.

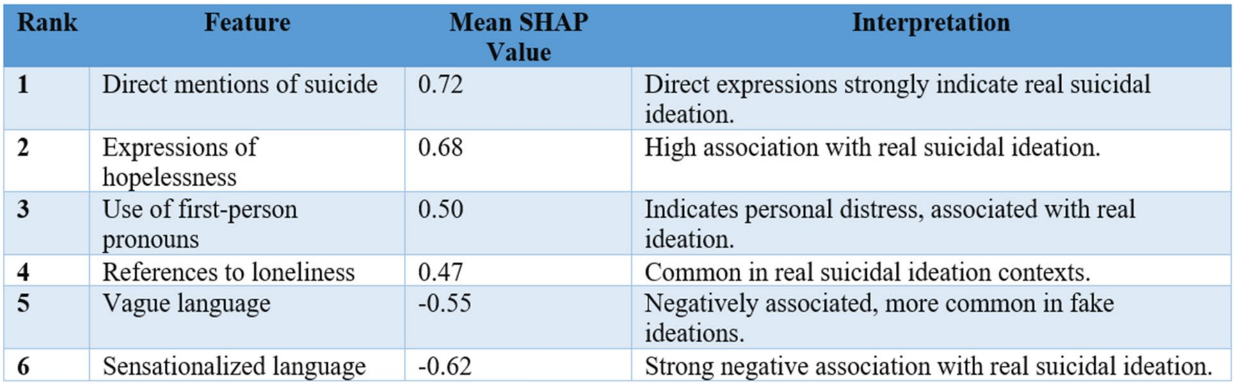
Cross validation

Ten equal portions of the dataset are randomly selected for tenfold cross-validation. Nine of these segments are used as the training set and the remaining one is used as the testing set for each iteration, Averaging the predicted results throughout the span of the 10 iterations yields the model's evaluation result[3,9].

It is critical to evaluate the test outcomes of the training set utilising learners from the prior layer during Meta-learner training. Predicting a trained learner inside the same training set can result in a label leak, disclosing personal information about dataset participants. To prevent label leaks during stacking, each training set receives an additional tenfold cross-validation. In this work, each Base-learner uses tenfold cross-validation to build a new feature, as shown in Fig. 5.

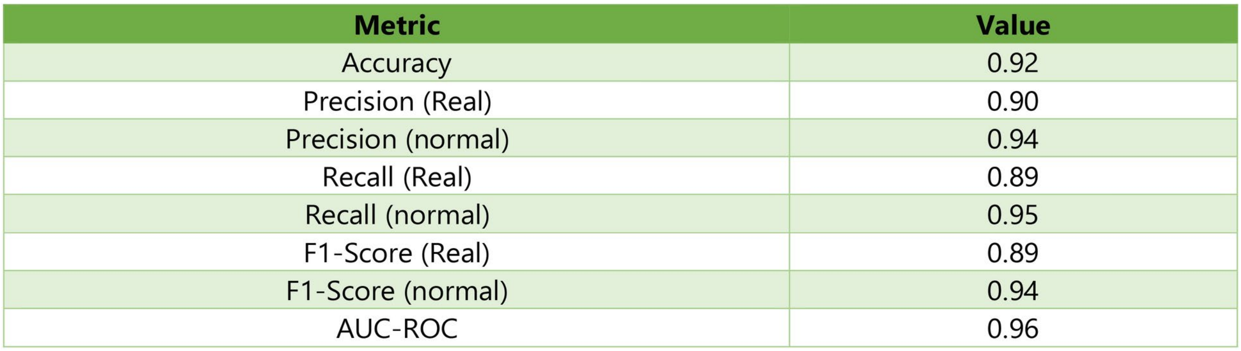
Model	Accuracy-mean	Standard-deviation	Macro-precision-mean	Standard-deviation	Macro-recall-mean	Standard-deviation	F1score-mean	Standard-deviation
SVM	90	0.02	90	0.04	91	0.04	91	0.04
LR	86%	0.04	88%	0.05	90%	0.04	89%	0.05
DT	84%	0.03	86%	0.04	88%	0.03	87%	0.04
GB	88%	0.04	89%	0.06	89%	0.04	88%	0.04
XSTM	92%	0.03	88%	0.03	91%	0.02	90%	0.04
Mamba	95%	0.02	89%	0.06	89%	0.03	88%	0.06
BERT (transformer)	89%	0.04	90%	0.04	90%	0.04	89%	0.03
Proposed (stacking-based ensemble method)	96%	0.02	93%	0.02	91%	0.03	92%	0.02

Table 10. tenfold cross-validation of the proposed ensemble model based on stacking model contrasted with other models.



Rank	Feature	Mean SHAP Value	Interpretation
1	Direct mentions of suicide	0.72	Direct expressions strongly indicate real suicidal ideation.
2	Expressions of hopelessness	0.68	High association with real suicidal ideation.
3	Use of first-person pronouns	0.50	Indicates personal distress, associated with real ideation.
4	References to loneliness	0.47	Common in real suicidal ideation contexts.
5	Vague language	-0.55	Negatively associated, more common in fake ideations.
6	Sensationalized language	-0.62	Strong negative association with real suicidal ideation.

Fig. 6. Feature importance analysis using SHAP values.



Metric	Value
Accuracy	0.92
Precision (Real)	0.90
Precision (normal)	0.94
Recall (Real)	0.89
Recall (normal)	0.95
F1-Score (Real)	0.89
F1-Score (normal)	0.94
AUC-ROC	0.96

Fig. 7. Predictive performance metrics results discussion.

Several classifiers are evaluated using tenfold cross-validation. The mean accuracy, standard deviation of accuracy, macro-precision, macro-recall, F-score, and standard deviation of F-score are all shown in Table 10. The results show that the proposed ensemble model based on stacking performed best, with the highest mean accuracy.

Results analysis for explainable AI (XAI) module

Let's discuss the results for the Explainable AI (XAI) module in terms of distinguishing genuine from non-suicidal thoughts in social media posts. The examination emphasizes the significance of features identified by SHAP values along, with the forecast accuracy measures derived from the XAI improved approach.

Results discussion

1. Feature Importance (Fig. 6): The study suggests that talking openly about thoughts of self harm expressing feelings of despair and using pronouns like "I" and "me" strongly suggest suicidal thoughts. On the hand vague language and exaggerated expressions are indicators of insincere suicidal ideation. This finding emphasizes how crucial it is to pay attention to language cues when determining the genuineness of expressions related to suicide.

Predictive Performance: In terms of performance (as shown in Fig. 7) the ensemble method, bolstered by XAI insights displays accuracy and precision in differentiating genuine from false suicidal thoughts. The impressive AUC ROC value indicates the models ability to discriminate effectively. It is worth noting that the model exhibits high precision, in detecting fabricated ideations likely attributed to recognizable patterns of exaggerated language that are simpler to categorize.

The data shown in these Figs gives us numbers to assess how well the Explainable AI system improves the accuracy of identifying thoughts of suicide in social media posts. The models ability to be both accurate and easy to understand marks a step, in using AI to support mental health.

2. The Generalizability of Experiment Results: Evaluating the generalizability of experimental findings is critical for assuring machine learning model stability and application in real-world contexts. Several factors influence the generalizability of the experimental findings in the context of Explainable AI-based suicidal and Non-Suicidal Ideations from Social Media Text with Enhanced Ensemble Technique utilizing the stacking classifier and four basic classifiers (SVM, decision tree, logistic regression, and gradient boosting).

Model Complexity: Overfitting, a phenomenon in which the model learns too much from the training data and fails to generalize to new data, can impede generalizability. To minimize overfitting, the stacking classifier and its basis classifiers are properly tuned. Furthermore, we employed Random Oversampling for skewed data handling (for example posts about suicidal thoughts).

Selection of Base Classifiers: The use of SVM, decision trees, logistic regression, and gradient boosting as base classifiers in this study demonstrates a purposeful desire to include a variety of algorithmic approaches. SVM is well-known for its capacity to handle complex relationships, decision trees provide interpretability, logistic regression is a standard linear model, and gradient boosting excels at increasing overall predictive power through repeated learning. This array of classifiers offers a layer of flexibility to the ensemble model, possibly accommodating different data distributions and patterns in the benchmark dataset.

Performance evaluation: The experimental results include metrics such as average accuracy, precision, recall, and F-score to assess generalizability. These metrics provide a thorough assessment of the model's performance across various suicidal ideations. Furthermore, the study will most likely include techniques such as tenfold cross-validation to assess the robustness of the suggested ensemble model against overfitting and variance concerns.

Data Representation: The reliability of the results depends on how the training data (X_Train) captures the language used to convey thoughts of suicide on social media platforms and, among different groups of people. Authors may consider using methods to expand the dataset to mitigate any prejudices.

Temporal Consistency: The way people talk about suicidal topics, on social media can change as time goes by. It's important to update the model, with data to make sure it stays accurate.

User-Centric Explanations: XAI explanations need to be customized for the intended audience. Although grasping the mechanisms of the model is beneficial, for researchers providing representations or summaries that are user friendly could be more beneficial for professionals, in intervention fields or social media platforms.

Conclusions and future work

This study investigated how Explainable AI (XAI) could be used to identify suicidal and non-suicidal thoughts expressed in social media posts. We utilized an method that involved a stacking classifier combining Support Vector Machine (SVM) Logistic Regression (LR) Gradient Boosting (GB) and Decision Tree (DT) models. The XAI element allowed us to grasp the reasoning, behind the models predictions. Proposed ensemble stacking classifiers have aggregated the predictions of various different models. In stacking classifier, it tries to enhance the accuracy of prediction, account for varied model power, balance bias and variance trade-offs, as well as it improves generalization by learning from individual model strengths. The stacking classifier is considered a super multi-layer. One or multiple models can be in each layer, and the results of the first layer are getting by the second layer of the model. Predictions from level first or base models are used to train the Meta classifier (Random Forest). The model receives textual data as input and can recognize and generate suicidal and non-suicidal thoughts. However, depending exclusively on text data for suicidal thoughts classification is insufficient, incorporating audio-visual, images, and even emoji's may enhance the study, giving for a more complete knowledge of a person's suicidal ideation features.

The proposed stacking classifier for suicidal ideation classification has produced improved results (suicidal = 95.5%, Non-Suicidal = 99%).

The application of XAI has shed light on the interactions, between language and emotional expression found in social media posts offering valuable perspectives on the indicators that differentiate authentic suicidal thoughts from non-suicidal material. This deeper comprehension demonstrates the potential of XAI to revolutionize the field of mental health-care assistance presenting a resource that acknowledges the subtleties of human communication while harnessing AI capabilities.

While we have made advancements, our exploration of understanding and reliable identification of signs of suicidal thoughts, on platforms is still ongoing. In discussing our model's generalizability, we acknowledge its performance may be confined to the specific social media platform and dataset used for training. Strategies to enhance generalizability, such as data augmentation or transfer learning, will be considered. We will address class imbalance issues, where a skew towards non-suicidal ideation may affect model performance, by exploring mitigation strategies like oversampling or cost-sensitive learning. While utilizing Explainable AI techniques, we recognize the limitations in fully understanding the decision-making processes of complex ensemble models, suggesting areas for further research in explainability. Regarding real-world application, we will discuss challenges such as interpreting sarcasm, slang, and addressing privacy concerns. The importance of human oversight and ethical considerations will be emphasized, especially in sensitive tasks like ideation detection.

For more improved results, we recommend future researchers to employ different feature selection techniques, such as mutual information, recursive feature removal and chi-square, to recognize the most suitable features and enhance the performance of the model. Additionally text representation techniques i.e. countvectorizer and tfidf, modern methods like contextual embedding (BERT) and word embedding (GloVe, Word2Vec,) may be investigated for better results. Moreover, SMOTE and other data balancing techniques may perfectly reflect the inequality distribution of negative and positive elements in actual environment, Furthermore, larger datasets or bootstrapping techniques can also be used to estimate the reliability and stability of the findings. Future studies will concentrate on aspects to enhance and broaden our approach. The primary focus will be on incorporating unique base classifiers to increase diversity within the ensemble, with the goal of increasing its resilience. Integrating cutting-edge algorithms and utilizing advances in machine learning techniques is critical for improving the ensemble model's forecasting accuracy. Another critical part is resolving the interpretability issue associated with ensemble approaches, which necessitates ways to clarify decision-making processes. We can explore the ways in which Explainable AI (XAI) can enhance teamwork, between people and artificial intelligence in intervention initiatives. This might include merging machine generated predictions with knowledge to create detailed evaluations. Furthermore, future research should concentrate on scaling issues, looking for ways to improve computational efficiency, particularly when dealing with increasingly vast and complicated datasets. Exploration of transfer learning approaches, which would allow the ensemble to use knowledge obtained from one suicidal dataset to boost performance on another, could be a promising direction. Adopting a cross-cultural viewpoint by verifying and modifying the ensemble model to other cultural situations would increase its usefulness and generalizability.

Data availability

Underlying data supporting the results can be provided by sending a request to the corresponding/submitting author.

Received: 28 March 2024; Accepted: 23 December 2024
Published online: 07 January 2025

References

1. Chatterjee, M., Kumar, P., Samanta, P. & Sarkar, D. Suicide ideation detection from online social media: A multi-modal feature based technique. *Int. J. Inf. Manag. Data Insights* **2**(2), 100103 (2022).
2. Nguyen, V. M., Nur, N., Stern, W., Mercer, T., Sen, C., & Bhattacharyya, S. (2023).
3. Liu, J., Shi, M. & Jiang, H. Detecting suicidal ideation in socialmedia: An ensemble method based on feature fusion. *Int. J. Environ. Res. Public Health* **19**(13), 8197–8197 (2022).
4. Malhotra, A. & Jindal, R. Xai transformer based approach for inter preting depressed and suicidal user behavior on online social networks. *Cogn. Syst. Res.* **84**, 101186 (2024).
5. Tadesse, M. M., Lin, H., Xu, B. & Yang, L. Detection of suicide ideation in social media forums using deep learning. *Algorithms* **1**, 7–7 (2019).
6. Chadha, A. & Kaushik, B. A hybrid deep learning model using grid search and cross-validation for effective classification and prediction of suicidal ideation from social network data. *N. Gener. Comput.* **40**(4), 889–914 (2022).
7. Aldhyani, T. H., Alsubari, S. N., Alshebami, A. S., Alkahtani, H. & Ahmed, Z. A. Detecting and analyzing suicidal ideation on socialmedia using deep learning and machine learning models. *Int. J. Environ. Res. Public Health* **19**(19), 12635 (2022).
8. Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L. & Thomson, J. An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *J. King Saud Univ. Comput. Inf. Sci.* **34**(10), 9564–9575 (2022).
9. Du, J. et al. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Med. Inf. Decis. Mak.* **18**, 77–87 (2018).
10. Choi, H. S. & Yang, J. Innovative use of self-attention-based ensemble deep learning for suicide risk detection in social media posts. *Appl. Sci.* **14**(2), 893–893 (2024).
11. Ji, S., Yu, C. P., Fung, S. F., Pan, S., & Long, G. Supervised learning for suicidal ideation detection in online user content. *Complexity* (2018).
12. Cheng, Q., Li, T. M., Kwok, C. L., Zhu, T. & Yip, P. S. Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study. *J. Med. Internet Res.* **19**(7), 243–243 (2017).
13. Choudhury, M. D., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 2098–2110 (2016).
14. Birjali, M., Beni-Hssane, A. & Erritali, M. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Proc. Comput. Sci.* **113**, 65–72 (2017).
15. Sawhney, R., Manchanda, P., Mathur, P., Shah, R., & Singh, R. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and socialmedia analysis*, pp.167–175 (2018).
16. Choudhury, M. D. & Kiciman, E. The language of social support in social media and its effect on suicidal ideation risk. *Proc. Int. AAAI Conf. Web Soc. Media* **11**, 32–41 (2017).
17. Roy, A. et al. A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ Digit. Med.* **3**(1), 78–78 (2020).
18. Alsulami, A. A. Enhancement of E-Learning student's performance based on ensemble techniques. *Electronics* **12**(6), 1508 (2023).
19. Moradi, M., Chen, Y., Du, X. & Seddon, J. M. Deep ensemble learning for automated non-advanced AMD classification using optimized retinal layer segmentation and SD-OCT scans. *Comput. Biol. Med.* **154**, 106512 (2023).
20. Nie, X. et al. Clustering ensemble in scRNA-seq data analysis: Methods, applications and challenges. *Comput. Biol. Med.* **159**, 106939 (2023).
21. [Online]. Available: https://www.kaggle.com/datasets/rvarun11/suicidal-ideation-reddit-dataset.

22. Kessler, R. C. et al. Clinical reappraisal of the composite international diagnostic interview screening scales (CIDI-SC) in the army study to assess risk and resilience in service members (Army STARRS). *Int. J. Methods Psychiatric Res.* **24**(3), 233–241. https://doi.org/10.1002/mpr.1471 (2015).
23. Allen, N. B., Nelson, B. W., Brent, D. & Auerbach, R. P. Short-term prediction of suicidal thoughts and behaviors in adolescents: Can recent developments in machine learning improve risk prediction?. *J. Affect. Disord.* **276**, 1142–1150. https://doi.org/10.1016/j.jad.2020.07.122 (2020).
24. Khan, G. A. S. et al. Personality classification from online text using machine learning approach. *Int. J. Adv. Comput. Sci. Appl.* **11**(3), 460–476 (2020).
25. Malhotra, A. & Jindal, R. Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Appl. Soft Comput.* **109**, 713 (2022).
26. Heckler, W. F., Carvalho, J. V. D. & Barbosa, J. L. V. Machine learning for suicidal ideation identification: A systematic literature review. *Comput. Hum. Behav.* **128**, 107095 (2022).
27. Ma, H. et al. Comprehensive learning strategy enhanced chaotic whale optimization for high-dimensional feature selection. *J. Bionic Eng.* **20**(6), 2973–3007 (2023).
28. Hou, D., Zhou, W., Zhang, Q., Zhang, K. & Fang, J. A comparative study of different variable selection methods based on numerical simulation and empirical analysis. *PeerJ Comput. Sci.* **9**, e1522 (2023).
29. Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4793–4813 (2020).
30. Komati, N. Suicide and depression detection (2021).

Acknowledgements

We sincerely thank the Information Security Research Group at King Abdulaziz University for their invaluable support and expertise, which significantly contributed to the success of this research.

Author contributions

All authors contributed equally.

Declarations

Competing interests

The authors declare no competing interests.

Informed consent

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Additional informed consent was obtained from all patients for whom identifying information is included in this article.

Human and animal rights

This study did not involve any experimental research on humans or animals; hence, an approval from an ethics committee was not applicable in this regard. The data collected from the online forums are publicly available data, and no personally identifiable information of the forum users were collected or used for this study.

Additional information

Correspondence and requests for materials should be addressed to M.Z.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.