Int J Comput Intell Syst

https://doi.org/10.1007/s44196-025-01123-9

Article in Press

Suicide Ideation Detection Using Social Media Data and Ensemble Machine Learning Model

Received: 27 June 2025

Accepted: 15 December 2025

Published online: 07 January 2026

Cite this article as: KINA E., Choi J., Ishaq A. *et al.* Suicide Ideation Detection Using Social Media Data and Ensemble Machine Learning Model. *Int J Comput Intell Syst* (2025). https://doi.org/10.1007/s44196-025-01123-9

Erol KINA, Jin-Ghoo Choi, Abid Ishaq, Rahman Shafique, Monica Gracia Villar, Eduardo Silva Alvarado, Isabel de la Torre Diez & Imran Ashraf

Suicide Ideation Detection Using Social Media Data and Ensemble Machine Learning Model

Suicide Ideation Detection using Social Media Data and Ensemble Machine Learning Model

EROL KINA[1], Jin-Ghoo Choi[2*], Abid Ishaq[3], Rahman Shafique[2], Monica Gracia Villar[4,5,6], Eduardo Silva Alvarado[4,7,8], Isabel de la Torre Diez[9], Imran Ashraf[2*]

[1]Ozalp Vocational School, Van Yuzuncu Yil University, Van, 65100, Turkey.
[2]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan, 38541, Republic of Korea.
[3]Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan.
[4]Universidad Europea del Atlantico, Isabel Torres 21, Santander, 39011, Spain.
[5]Universidad Internacional Iberoamericana Arecibo, Puerto Rico, 00613, USA.
[6]Universidade Internacional do Cuanza, Cuito, Angola.
[7]Universidad Internacional Iberoamericana, Campeche, 24560, Mexico.
[8]Universidad de La Romana, La Romana, Republica Dominicana.
[9]Department of Signal Theory and Communications and Telematic Engineering, University of Valladolid, Paseo de Belen, 15, Valladolid, 47011, Spain.

*Corresponding author(s). E-mail(s): jchoi@yu.ac.kr; ashrafimran@live.com;
Contributing authors: erolkina@yyu.edu.tr; abid.ishaq@iub.edu.pk; rahmanshafique47@gmail.com; monica.gracia@uneatlantico.es; eduardo.silva@uneatlantico.es; isator@tel.uva.es;

Abstract

Identifying the emotional state of individuals has useful applications, particularly to reduce the risk of suicide. Users' thoughts on social media platforms can be used to find cues on the emotional state of individuals. Clinical approaches to

suicide ideation detection primarily rely on evaluation by psychologists, medical experts, etc., which is time-consuming and requires medical expertise. Machine learning approaches have shown potential in automating suicide detection. In this regard, this study presents a soft voting ensemble model (SVEM) by leveraging random forest, logistic regression, and stochastic gradient descent classifiers using soft voting. In addition, for the robust training of SVEM, a hybrid feature engineering approach is proposed that combines term frequency-inverse document frequency and the bag of words. For experimental evaluation, "Suicide Watch" and "Depression" subreddits on the Reddit platform are used. Results indicate that the proposed SVEM model achieves an accuracy of 94%, better than existing approaches. The model also shows robust performance concerning precision, recall, and F1, each with a 0.93 score. ERT and deep learning models are also used, and performance comparison with these models indicates better performance of the SVEM model. Gated recurrent unit, long short-term memory, and recurrent neural network have an accuracy of 92% while the convolutional neural network obtains an accuracy of 91%. SVEM's computational complexity is also low compared to deep learning models. Further, this study highlights the importance of explainability in healthcare applications such as suicidal ideation detection, where the use of LIME provides valuable insights into the contribution of different features. In addition, k-fold cross-validation further validates the performance of the proposed approach.

Keywords: Suicide ideation; machine learning; feature extraction; ensemble learning; feature fusion

1 Introduction

A person dies by suicide approximately every forty seconds, making it a critical global health concern, as per the World Health Organization (WHO) [1]. The WHO report highlights that around 16 million people attempt suicide every year, with nearly 800,000 fatalities [2]. However, due to significant underreporting, around one million people annually are believed to die of suicide [3]. Statistics further suggest that young people, including women, are among the top victims of suicide. It is important to understand that suicide is not a binary outcome but rather a process with distinct stages. A well-known book on the subject [4] suggests that it follows a specific pattern, starting with suicidal ideation, followed by a suicide attempt, and ultimately leading to completed suicide. While not all instances of suicidal ideation result in suicide, they present a substantial risk for individuals who may then proceed to attempt suicide.

The fourth leading cause of death among young people between 15 and 29 years of age is suicide. Additionally, it is believed that suicide attempts are 20 times that of deaths caused by it [5]. Environmental factors, health factors, and personal abuse are among the top three factors being the reason for suicides [6, 7]. Physical illnesses, domestic violence, substance abuse, bullying, mental disorders, significant life stressors, relationship issues, and financial problems are a few of the other risk factors of suicide. Given the complicated nature of the issue, it is not possible to rely on a single risk factor to accurately predict suicide [8]. Although depression is strongly

associated with suicide, a diagnosis of depression alone has limited predictive ability. Furthermore, the COVID-19 pandemic has exacerbated the issue of suicide [9]. The measures implemented to control the virus, such as social isolation, have been linked to an increased threat of suicide.

Individuals are categorized into suicide ideators and suicide attempters on the basis of associated suicide risk factors [10]. Suicidal ideation encompasses a range of behaviors and thoughts, from being anxious about death to aggressively attempting suicide. Suicidal ideation can either be active or passive. Active suicidal ideation involves the planning and intention of a suicide attempt, while passive suicidal ideation involves views about suicide and a craving to be dead [11]. Although passive suicidal ideation is generally considered to be less risky, both forms should be sensibly evaluated by psychologists, as the passive form can quickly translate into active ideation [12]. This transformation may occur when an individual's situation or health deteriorates.

Timely detection of suicidal ideation remains a challenge, and there is a need for more effective methods to detect and intervene before they engage in suicidal behavior [13]. However, several obstacles hinder suicide prevention efforts. These challenges encompass (1) social stigma surrounding mental health, (2) limited availability of professional support, and (3) insufficiently trained clinics related to suicidal management [14]. This amalgamation of factors gives rise to an additional challenge in the form of fragmented professional care, resulting in significant intervals between mental health assessments.

1.1 Problem Statement

Due to the rise and rapid growth of social media platforms, the trend of individuals openly discussing suicidal thoughts within online communities has become common [15]. Content uploaded on such platforms can potentially be used to detect suicidal intentions. This is especially common among adolescents, who are more open to sharing such thoughts publicly, and ask for advice on suicidal methods in online groups [16]. The secrecy afforded by social media communication enables individuals to openly explain their anxieties and social or mental pressures they encounter publicly. This user-generated social media online content offers a new avenue for the early detection and prevention of suicide.

1.2 Research Gap

The use of social media for detecting suicidal ideation has the potential to reduce suicide in society. It can help health professionals and psychologists quickly identify at-risk individuals and intervene in a timely manner. This idea has already been explored using real-world datasets, receiving positive feedback from psychiatrists [17]. Chiang et al. offered an early warning system for suicidal tendencies using social networking platforms, thereby enabling timely interventions by psychologists [18]. Integration of ML and NLP can help develop automated systems for the detection of suicidal ideation, addressing some of the challenges in early identification. However, the current systems lack the desired level of performance and accuracy for these platforms and require

further research. Furthermore, it is very important to establish clear explainability criteria for the decision-making process of AI-based approaches, which has been lacking in previous studies.

1.3 Contributions

It is important to recognize that forecasting mental illnesses and suicide ideation among individuals using NLP and ML algorithms should be considered an initial step in spreading awareness about one's mental state, rather than portraying conclusive assumptions. It is crucial to exercise caution and rely on professional mental health assessments for accurate diagnosis and intervention. With that in mind, this research's contributions are as follows.

- **Development of a lightweight ensemble learning model:** An ensemble model is designed to provide enhanced accuracy for suicide ideation detection by combining random forest (RF), logistic regression (LR), and stochastic gradient descent classifier (SGDC). The proposed model offers better performance with lower computational complexity.
- **Proposal of hybrid feature engineering:** A hybrid feature set is proposed to fully utilize the potential of term frequency-inverse document frequency (TF-IDF) and bag of words (BoW). Combining these features helps to reduce the uncertainty that may arise due to using a single feature engineering approach. Furthermore, confidence, reliability, and information quality can be improved using hybrid features. Separate experiments are conducted to analyze the effectiveness of various feature engineering approaches: TF-IDF, global vectors (GloVe), and BoW.
- **Extensive experimental evaluation:** To verify performance of SVM for suicidal tendencies identification, various ML and deep learning (DL) models are also used including LR, SGDC, Naive Bayes (NB), AdaBoost (ADA), long short-term memory (LSTM), convolutional neural network (CNN), gated recurrent unit (GRU), and recurrent neural network (RNN). In addition, k-fold cross-validation and performance comparison with state-of-the-art methods are also carried out to further validate the performance of the proposed approach.
- **Use of LIME explainability:** To provide insights into the decision-making rationale of ML models, the local interpretable model-agnostic explanation (LIME) approach is also applied.

The remaining portion of this study is structured into different segments. Section 2 covers the existing research conducted on identifying and preventing suicidal ideations. Section 3 outlines the proposed approach, comprising several subsections: dataset, feature engineering techniques, data preprocessing, and ML models used in this study. Section 4 is about results and analysis. Lastly, Section 5 presents the conclusion and explores potential future directions for this study.

2 Related Work

The use of ML and DL has been reported to improve performance in several applications [19–21]. Similarly, feature engineering approaches are important for better model

training, thereby leading to better outcomes [22]. The rising trend of suicide in recent years has sparked the interest of numerous researchers, leading to a focus on suicide detection. In various NLP projects, text extracted from Reddit, TikTok, Instagram, Facebook, and Twitter social media platforms [23], and similar forums have established effectiveness in ascertaining users having suicidal thoughts, mental illness, and social distress. This approach enables the determination of users' suicidal ideation based on text extraction techniques from their posts. ML techniques, particularly NLP and supervised learning methods, have been widely employed in several applications [24–26], particularly for the prediction of suicidal ideation [27].

An ensemble learning model based on DL is proposed by Renjith et al. [28] for suicide ideation detection from posts on different social media platforms. The authors used the ML models individually and in the ensemble. They proposed an ensemble learning model, LSTM-attention-CNN, for the detection of emotions from social media posts. Results of the study show that the proposed ensemble model achieved an accuracy of 90.3%, which is much better than the baseline models. For the automatic detection of suicidal ideation, an LSTM-based system was proposed by Deepaj et al. [29]. The authors used DL and ML models for the efficient detection of suicide ideation using Twitter data. To extract the tweets, they used the publicly available application programming interface (API). Results show that the LSTM achieved the highest accuracy of 90.8% for discriminating suicidal tweets from non-suicidal tweets.

Chatterjee et al. [30] conducted a study on the detection of suicidal ideation by analyzing online user-generated content. The study involved each user providing two inputs to the model. The first input utilized a set of linguistic attributes to represent the individual user's tweets. The second input involved assessing the user's personality through a word-by-word emotional and sentiment analysis of their tweets. The study encompassed various steps, including data preparation, feature extraction, analysis, fusion, and classification. Additionally, the SVM model combined with Trigram+LDA+TS+TF-IDF+TA+EA+SA features obtained an accuracy of 87%. A code-mixed Hindi-English dataset is used for suicide ideation detection by Agarwal and Dhingra [31]. The authors used DL and ML models for this purpose. Results indicate that the attention-BiLSTM achieved the highest accuracy score of 90.66%. The authors also used the bidirectional encoder representation from the Transformer (BERT) approach, which showed an accuracy of 98.54%.

An ML-based approach was proposed by Bandari et al. [32]. The authors used a single ML model, SGDC, with different training and testing set splits. Results of the study show that SGDC with 25% training data achieved a classification accuracy of 91.6%. The study [33] used the benchmark dataset for the detection of suicide ideation in online user content. They used ML and DL models with different feature fusion techniques. Experimental results show that the RF achieved an accuracy of 96.38% on the pre-processed features.

Nordin et al. [34] proposed a predictive model for the prediction of suicide attempts. The study used eight ML models, such as LR, DT, SVM, NB, KNN, RF, bagging, and voting, for this purpose. They used a three-fold cross-validation technique for the feature extraction. Results of the study revealed that the ensemble learning models, bagging and voting, achieved the highest accuracy of 92%. Similarly, [35] used the DL

models for suicide ideation detection. The study proposed a model for audio and text classification. Results reveal that the deep neural network (DNN) outperformed other models for the emotion classification and achieved an accuracy of 88%, and for the text classification, ROBERTA achieved an accuracy of 98.81%.

Tadesse et al. [36] proposed a DL architecture for suicide ideation through social media posts. For this purpose, the authors used ML and DL techniques with different feature engineering methods. The result of the study explores that the ensemble DL model LSTM-CNN achieved a 93.8% accuracy with word2Vec features.

The study [37] analyzes tweets related to COVID gathered using the Twitter API. The collected tweets are from diverse domains such as health, politics, education, etc. The NRC emotion lexicon is used to label tweets. The focus is to classify emotions into various categories, like fear, joy, trust, etc. The study presented an emotion care framework and an online platform to assess people's mental state during COVID and provided superior performance concerning the accurate recognition of various emotions. Similarly, [38] focused on detecting abusive content from social media platforms in the low-resource Tamil language. The research particularly investigated the use of various DL models to classify text written in the Malayalam and Tamil languages. Results indicate that BERT-based models tend to show better results, with IndicBERT showing the best results with 0.86 and 0.77 F1 scores for Malayalam and Tamil languages, respectively. The study [39] carries out the important task of emotional response detection due to misinformation and fake news shared on social media platforms. For better accuracy, an ensemble model based on hard voting is designed. Compared to other models used in the study, the proposed ensemble model exhibited a superior accuracy of 93.48%.

Recently, several studies used large language models (LLMs) for suicide ideation detection using data from social media. For example, [40] used triple word embedding models for suicidal thoughts detection. Besides the CNN and BiLSTM, the study also investigated BERT and generative pre-training transformer (GPT) models for this task. The proposed GPT model showed much better results with 97.69% accuracy. Similarly, the authors [41] investigate pre-trained language models to detect the mental stage of users. Four BERT models of Hugging Face were utilized in the study to analyze disorder symptoms. Of the use models, the BERT-based-uncased (BBU) model showed the best performance with 97.00% accuracy.

The study [42] utilized Weibo posts to analyze thoughts that might lead to suicide. The C4.5 model was developed for experiments, and the top 20 features were selected using three feature engineering approaches. Using only the top 6 features, the C4.5 model showed an AUC of 0.97. The authors [43] perform validation on the proposed suicide detection algorithms using data from two cohort studies. The data was taken from X (Twitter), as well as from recruited independents who participated in the survey. Experimental results indicate that using the proposed algorithm, the association between stress and suicide can be established. In addition, the suicidal trajectory caused by the use of social media can also be established. For the research works discussed in this section, a comparative overview is presented in Table 1, where accuracy (Acc.), ensemble learning (EL), explainable AI (XAI), feature fusion (FF), and complexity (Com) are summarized; ✓ indicates presence, and × indicates absence.

Table 1: Analytical comparison of related work.

Ref.	Model(s)	Dataset (type/size)	Acc. (%)	EL	XAI	FF	Com
[28]	CNN, LSTM, SVM	Reddit (11k posts)	90.3	✓	×	×	Medium
[29]	LR, DT, CNN, LSTM	Twitter (16.8k tweets)	90.8	×	×	×	Medium
[30]	LR, RF, SVM, XGB	Twitter (188k tweets)	87.0	×	×	✓	Medium
[31]	SVM, CNN, BiLSTM	Hindi-English (6.5k posts)	90.7	×	×	×	High
[32]	SGDC	Twitter (9.1k tweets)	91.6	×	×	×	Low
[33]	GBDT, SVM, LSTM, RF	Reddit+Twitter (3.5k posts)	96.4	×	×	✓	High
[34]	LR, SVM, RF, Bagging	Clinical (75 patients)	92.0	✓	×	×	Medium
[35]	RF, CNN, LSTM, DNN	Weibo (40k posts)	98.8	×	×	✓	High
[36]	RF, SVM, CNN, LSTM	Reddit+Twitter (7k posts)	93.8	×	×	✓	High
Our	SVEM	Suicide and Depression ($\tilde{1}$32k)	94.10%	✓	✓	✓	Low

As shown in Table 1, most existing studies achieve strong performance but lack either ensemble strategies, XAI, or effective feature fusion. Only a few works integrate multiple approaches simultaneously, and many rely on complex deep learning architectures, which limit scalability and interpretability. Moreover, explainability is almost entirely absent in prior studies, despite its importance in sensitive domains like mental health. These gaps highlight the need for a lightweight, interpretable ensemble framework that combines hybrid feature fusion with high accuracy.

3 Materials and Methods

This section discusses the dataset description, data preprocessing techniques, feature engineering methods, the description of ML models for suicide ideation detection, and the workflow of the proposed methodology. Figure 1 shows the workflow diagram of the adopted methodology.

Initially, we obtained a dataset containing posts about depression from the publicly available Kaggle repository. This dataset is categorized into two classes and consists of social media posts in text form, as outlined in Section 3.1. Following dataset extraction, we carried out data preprocessing to eliminate irrelevant raw content present in the posts. This step was necessary, as the raw data would not contribute meaningfully to training the machine learning models. In the text preprocessing phase, we applied several techniques, including stopword removal, punctuation removal, conversion to lowercase, stemming, and lemmatization. After completing preprocessing, the dataset was split into training and testing sets using a 70:30 ratio. Feature extraction techniques were then applied, as the textual data cannot be directly fed into the learning models. To address this, we employed methods that convert text into numerical representations. Additionally, hybrid features were generated by combining Term Frequency–Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) approaches. This combination leverages both frequency-based and weighted features, yielding more meaningful and discriminative representations. Subsequently, we trained several machine learning models and further introduced an ensemble model comprising three distinct classifiers, as detailed in Section 3.6. Finally, the methodology was evaluated using accuracy, precision, recall, F1 score, and confusion matrix analysis.

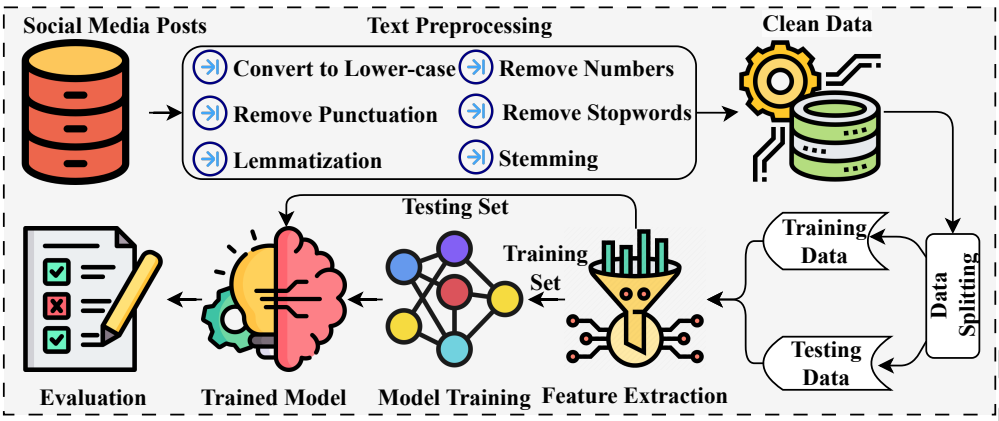

Fig. 1: Workflow diagram of the proposed methodology.

3.1 Dataset Description

The dataset used in this study was obtained from the Kaggle repository, a reputable data source. The dataset, "Suicide and Depression Detection," is publicly accessible on the website[1] [44]. It consists of posts extracted from the "SuicideWatch" and "depression" subreddits on the Reddit platform. The posts were collected using the Pushshift API. All posts from the "SuicideWatch" subreddit between December 16, 2008, and January 2, 2021, were included, while posts from the "depression" subreddit were collected from January 1, 2009, to January 2, 2021. The dataset contains 232074 instances and consists of two features: "text" and "class" as shown in Table 2.

Table 2: Dataset attributes.

Attribute.	Description
text	This attribute contains social media text posted by users.
class	The "class" feature serves as the target class and has two categories: suicide and non-suicide.

The "SuicideWatch" subreddit posts were categorized as 'SuicideWatch' to indicate their association with suicide-related content. Additionally, posts from other subreddits such as 'depression' and 'teenagers' were collected from their respective subreddits. Table 3 presents some examples of instances from the dataset.

[1] https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch/data

Table 3: A few sample instances from the dataset.

Text	Class
I am getting pretty closeI just dont see any other option at this point	suicide
Would be fun to make a group chats with random people Post is in the title	non-suicide
I don't know what to do anymoreI'm 17, I've been depressed for 8 months and had thoughts of killing ...	suicide
I have a question for you all. Where the fuck is male snoo in the banner?	non-suicide
Oct 22 2016Jag vill dö.	suicide
Anybody know where to download the emoji keyboard? I deleted and now i need to use it for okbr	non-suicide

3.2 Data Visualization

Figure 2 shows the distributions of the ratio of both classes in the dataset. It reveals that there is an equal distribution between the suicide and non-suicide classes, with both classes accounting for 50% of the instances.

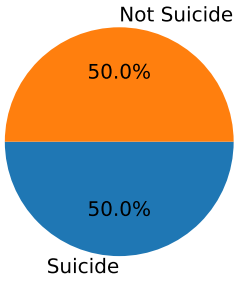

Fig. 2: Class distribution in dataset.

Based on the information presented in Table 4, we can ascertain that the dataset consists of a total of 232,074 instances. Among these instances, 116,037 are categorized as suicide instances, while the remaining 116,037 instances are classified as non-suicide instances. This balanced distribution indicates that the dataset is evenly divided between the two classes.

Table 4: Class-wise distribution in the dataset.

No.	Category	No of instances
1	suicide	116,037
2	non-suicide	116,037

3.3 Data Preprocessing

Transforming unorganized data into a structured format is a crucial step in the preprocessing of input data. It enhances the quality of the input data to effectively identify the trends and extract relevant features by the model [45]. The text related to suicide ideation exhibits a mixture of lower-case and upper-case letters and words, containing punctuation and stopwords, impeding the model's learning capability. Table 5 presents two sample texts from the dataset, providing an illustration of its structure.

Table 5: Sample text from the dataset.

No.	Posts Text
1	The days just keep coming, Another day another struggle.
2	I have 27 pages of sheet music, So I auditioned as a singer for this music school.

Figure 3 shows the steps performed during the preprocessing of data. The preprocessing procedures for this study consist of multiple stages, including text case conversion, punctuation elimination, stopwords removal, elimination of null and numeric values, and stemming. The order in which these preprocessing steps are carried out is depicted in Figure 3, and a concise explanation is provided for each step.

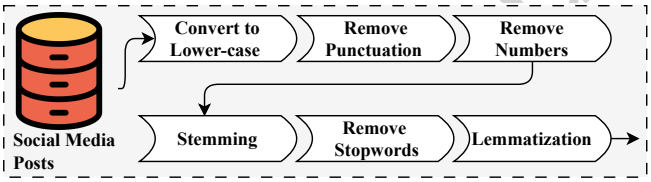

Fig. 3: Flow of preprocessing of the data.

3.3.1 Lowercase Conversion

Probabilistic ML models exhibit case sensitivity, meaning that words like 'Energy' and 'energy' are considered distinct entities by the model. To address this, the initial preprocessing step is to change social media posts' text to lowercase. This ensures that words with different cases are treated as the same word. Table 6 displays a representation of sample social media posts after the conversion to lowercase has been applied.

Table 6: Sample posts after conversion to lowercase.

Suicide post	Converted to lowercase
The days just keep coming, Another day another struggle.	The days just keep coming, another day another struggle.
I have 27 pages of sheet music, So I auditioned as a singer for this music school.	I have 27 pages of sheet music, so I auditioned as a singer for this music school.

3.3.2 Punctuation removal

In the second step, various punctuation marks such as $, %, #,), !, }, & (. ?, ” are eliminated from the reviews. The removal of punctuation is necessary in this study as it does not provide meaningful contributions. Additionally, keeping punctuation intact can hinder the algorithm's ability to differentiate between punctuation marks and other characters. Table 7 presents the resulting sample posts after the removal of punctuation.

Table 7: Sample posts after punctuation removal.

Suicide post	Punctuation removal
the days just keep coming, another day another struggle.	the days just keep coming another day another struggle
i have 27 pages of sheet music, so i auditioned as a singer for this music school.	i have 27 pages of sheet music so i auditioned as a singer for this music school

3.3.3 Number and Null Values Removal

In this step, null values and numeric values are eliminated from the suicide posts. Since the focus is on textual data and reviews, digits are not relevant and therefore removed during preprocessing. This principle is applicable to null values, as they do not enhance the model's performance. Furthermore, numeric and null values do not impact the sentiment scrutiny of the text. Table 8 illustrates the resulting sample posts after the removal of null and numeric values.

Table 8: Sample posts after the number and null values removal.

Suicide post	Removing number and null values
the days just keep coming another day another struggle	the days just keep coming another day another struggle
i have 27 pages of sheet music so i auditioned as a singer for this music school	i have pages of sheet music so i auditioned as a singer for this music school

3.3.4 Stopwords Removal

After numeric and null values are eliminated, the reviews undergo the removal of stopwords. This step is a crucial part of preprocessing as it eliminates irrelevant data for text analysis. Stopwords are words that hold no significance for predictive modeling. The output of sample reviews after the removal of stopwords can be observed in Table 9.

Table 9: Sample posts after stopwords removal.

Suicide post	Removing stopwords
the days just keep coming another day another struggle	days just keep coming another day another struggle
i have pages of sheet music so i auditioned as a singer for this music school	pages sheet music auditioned singer music school

3.3.5 Stemming

Following the stopwords removal, stemming is carried out on the suicide posts. Stemming involves converting words into their root form, aiming to enhance the performance of a model. For instance, the word 'play' may appear in various forms, such as 'played' or 'playing,' which would be considered distinct words by ML models. Therefore, performing stemming is essential to convert them into their root form. Table 10 displays the output of sample suicide posts after stemming has been carried out.

Table 10: Sample posts after stemming

Suicide post	Stemming
days just keep coming another day another struggle	day just keep come another day another struggle
pages sheet music auditioned singer music school	page sheet music audition singer music school

3.4 Feature Engineering

Feature engineering involves the exploration of valuable data features or the creation of new features from existing ones to effectively train ML algorithms. Its purpose is to enhance the efficiency of these algorithms. In the realm of ML, there is a proverb called "garbage in, garbage out" which emphasizes that nonsensical input data will lead to meaningless output [46]. Conversely, utilizing information-rich data yields more favorable results. Therefore, feature engineering plays a crucial role in extracting valuable insights from raw data, thereby improving the reliability and accuracy of learning algorithms. In the proposed methodology, four feature engineering techniques are employed, namely BoW, TF-IDF, GloVe, and feature fusion (BoW+TF-IDF).

3.4.1 Bag of Words

The technique known as BoW is employed to derive features from textual data. BoW is not only straightforward to implement and comprehend but also represents the most basic approach for extracting features from text [47]. BoW proves highly advantageous and valuable in tasks such as language modeling and text classification. To implement BoW, the 'CountVectorizer' library is utilized. CountVectorizer computes the frequency of words and generates a sparse matrix of word occurrences. BoW essentially functions as a collection of words or features, with each feature labeled to indicate its corresponding frequency of occurrence.

3.4.2 Term Frequency-Inverse Document Frequency

TF measures the frequency of a word within a document, and IDF quantifies the rarity of a word across a corpus of documents. TF-IDF, a statistical analysis technique, combines both measures to determine the relevance of words in a given list or corpus [48]. The TF-IDF value increases as a word appears more frequently, but is normalized by its occurrence in the text. TF can be represented as

$$F(t) = \frac{No.\ of\ \ times\ \ t\ \ appears\ \ in\ \ a\ \ document}{Total\ \ no.\ of\ \ terms\ \ in\ \ the\ \ document} \tag{1}$$

IDF can be represented as

$$IDF(t) = \log \frac{Total\ \ no.\ \ of\ \ document}{No.\ \ of\ \ documents\ \ with\ \ term\ \ t\ \ in\ \ it} \tag{2}$$

TF-IDF is then calculated using both TF and IDF, using

$$TF - IDF = TF_{(t,d)} * \log\left(\frac{N}{Df}\right) \tag{3}$$

where $TF_{t,d}$ represents the frequency of term t in document d, N is the total number of documents, and Df is the number of documents containing term t.

3.4.3 Global Vectors

GloVe (Global Vectors) for word representation is a method used to generate word embedding by establishing relationships between words. It accomplishes this by combining global co-occurrence matrices containing word pair frequency appearing together [49]. Through the co-occurrence matrix of a quantity, words are segregated into groups based on similarity and dissimilarity among them. The GloVe model weights statistical data in a word-word co-occurrence matrix from nonzero elements exclusively.

3.4.4 Hybrid Features

In order to enhance the performance of ML models, we suggest combining the features of BoW and TF-IDF [50]. This concatenation, as illustrated in Figure 4, proves beneficial for learning models, leading to improved accuracy.

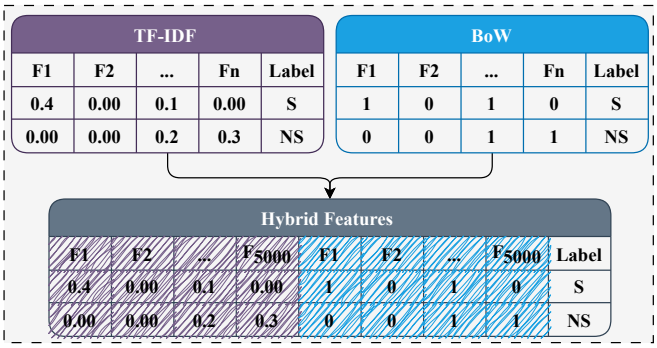

Fig. 4: Feature fusion process.

3.5 Machine Learning Classifiers

In this study, supervised ML classifiers are employed to predict target variables based on the available data. The implementation of these ML classifiers is carried out using the Python programming language, by using the 'sci-kit-learn' Python library. The classifiers are trained from the training set on known data samples and then validated with the dataset, which is unknown and unique to the classifier, to evaluate their learning competencies and performance. Table 11 shows the hyperparameter settings of the ML models along with the tuning values. We use a grid search approach to run the models with different hyperparameter settings and then select the best values for our work.

Table 11: Hyperparameter settings of supervised ML classifiers.

Model	Hyperparameter settings	Hyperparameter Tuning
LR	random_state= 1000, solver= 'liblinear', multi_class= 'ovr', C= 1.0	solver= ['liblinear', 'sag', 'saga'], C= [1.0 to 5.0]
RF	n_estimators= 300, random_state= 2, max_depth= 300	n_estimators= [50 to 500], max_depth= [50 to 500]
SGDC	max_iter= 100, tol= 1e-3,	– – –
NB	Default settings	– – –
ADA	n_estimators= 100, random_state= 5, learning_rate= 0.2	n_estimators= [50 to 500], learning_rate= [0.01, 0.1, 0.2, 0.5, 0.8]

3.5.1 Logistic Regression

LR is a statistical ML classifier that maps discrete target variables to input features using a sigmoid function, providing probability approximations. This function confines probabilities within the distinct target variable range, making it suitable for grouping tasks. LR is adept at handling both linear and non-linear datasets, making it versatile for classification and prediction [51]. Widely used for modeling Boolean data, LR multiplies weight values with input values. Known for detecting defaulters effectively,

LR is renowned for its versatility in handling extensive classification problems and relies on fewer assumptions than other algorithms.

In this study, LR uses the 'liblinear' solver, known for computational efficiency, ideal for large datasets. The 'multi_class' parameter is set to 'ovr' (one-vs-rest) for binary classification tasks. 'C,' the inverse regularization parameter, is set to 0.1, controlling regularization strength to mitigate overfitting as shown in Table 11. A smaller 'C' value strengthens regularization, enhancing generalization capabilities.

3.5.2 Random Forest

RF is a versatile ensemble model for classification and regression, using bagging to generate multiple trees with majority voting for predictions. Increased tree count in RF generally enhances prediction accuracy [52]. RF effectively mitigates overfitting by employing the bootstrap sampling method.

The hyperparameters for the RF are shown in Table 11. For this study, RF utilizes 300 decision trees, determined by the 'n_estimators' parameter. To control tree complexity and enhance model performance, the 'max_depth' parameter is set to 300, restricting the maximum depth of each tree.

3.5.3 Stochastic Gradient Descent Classifier

SGDC is an iterative optimization algorithm commonly used for training ML models. It determines function coefficients and parameter values with convex loss functions [53]. The coefficients are individually trained for each instance, making them suitable for large-scale datasets and significantly reducing computational costs, especially in high-dimensional optimization problems. This approach prioritizes quicker repetitions, albeit with a slightly slower convergence rate.

SGDC employs tuned hyperparameters, including 'tol' set to 1e-3, serving as the convergence benchmark as shown in Table 11. Training stops when successive epochs, determined by this parameter, are reached. The 'max_iter' parameter, set to 100, defines the maximum number of training iterations.

3.5.4 Adaptive Boosting

ADA is an iterative ensemble technique that builds robust learners by combining multiple weak learners, aiming to minimize errors through iterative training on weighted examples. As a meta-estimator, ADA initially fits a classifier on the original dataset [54]. Subsequently, it trains additional copies of the classifier, adjusting weights for poorly classified instances. This enables subsequent classifiers to concentrate more on complex scenarios.

This study employs ADA with tuned hyperparameters, including n_estimators and random_state (Table 11), aiming for higher accuracy. The algorithm combines 100 weak learners, specified by n_estimators, to achieve the final forecast. The boosting process concludes upon reaching the maximum estimators or achieving a perfect fit. The random_state parameter, set to 5, controls the randomness in sample selection during training, ensuring limited randomness for each boosting iteration.

3.5.5 Gaussian Naive Bayes

The GNB model is based on Bayes' theorem. When implementing it, the classifier makes certain assumptions, such as the independence of all features included in the model [55]. It is commonly used for classifying objects with data that follows a normal distribution. Because of these characteristics, it is often referred to as the Gaussian Naive Bayes classifier. We use it with the default hyperparameter settings.

3.6 Proposed Soft Voting Ensemble Model

SVEM, the proposed ensemble model, uses a soft voting scheme to perform classification by merging the output of various base models. It combines the results from individual classifiers to arrive at the final prediction. This study combines RF, SGDC, and LR under soft voting criteria. Each model within our ensemble serves a distinct purpose, which clarifies the motivation behind our model selection and its integration into the ensemble:

- RF is chosen for its capability to handle high-dimensional data, capture complex relationships between features, and effectively handle noisy data. Its ability to mitigate overfitting while maintaining robustness made it an essential component of our ensemble [56].
- LR and SGDC are incorporated due to their interpretability, efficiency in handling large datasets, and ability to provide probabilistic predictions [57]. These models serve as complementary components, contributing to the diversity of our ensemble.

The rationale behind this selection is to combine models with varying strengths and weaknesses to create an ensemble that compensates for individual model limitations while capitalizing on their collective strengths. SVEM can be expressed as:

$$\widehat{p} = argmax\{\sum_{i}^{n} RF_i, \sum_{i}^{n} LR_i, \sum_{i}^{n} SGDC_i\}. \tag{4}$$

where $\left(\sum_{i}^{n} RF_i, \sum_{i}^{n} LR_i, and \sum_{i}^{n} SGDC_i\right)$ represent n prediction probabilities for a sample. Each prediction probability for a sample is subjected to the soft voting criteria, as shown in Figure 5.

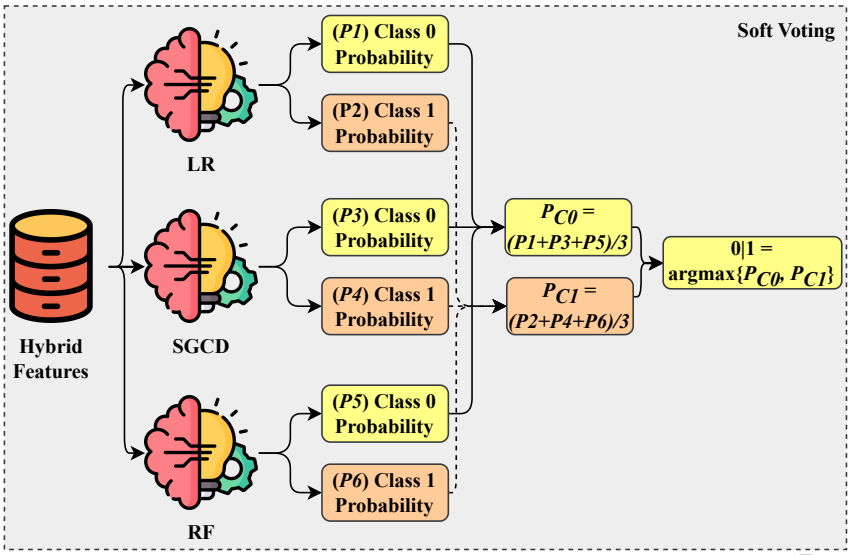

Fig. 5: Voting mechanism in the proposed ensemble model.

The soft voting criteria consider the combined probabilities from RF, LR, and SGDC classifiers and use a specific method to determine the final prediction for the target class. Figure 5 visualizes the process of aggregating the probabilities and applying the soft voting criteria to arrive at the ultimate prediction. Soft voting criteria can be defined mathematically as:

$$\begin{aligned} \text{Avg PC1} &= \frac{PC1_{M1} + PC1_{M2} + \ldots + PC1_{Mn}}{n} \\ \text{Avg PC2} &= \frac{PC2_{M1} + PC2_{M2} + \ldots + PC2_{Mn}}{n} \\ &\vdots \\ \text{Avg PCm} &= \frac{PCm_{M1} + PCm_{M2} + \ldots + PCm_{Mn}}{n} \end{aligned} \tag{5}$$

where (Avg PC1, Avg PC2, ..., Avg PCm) are the average probabilities for each class, and m is the number of classes. Here, $(PC1_{M1} + PC1_{M2} + \ldots + PC1_{Mn})$ represents the probabilities generated by n models for class 1, where n is the number of models.

Algorithm 1 SVEM algorithm for suicide ideation detection.

Input: Corpus - Text posts (i)
Output: Suicide (Class 1) or Non-Suicide (Class 0)

$LR_t \longleftarrow$ Trained LR
$RF_t \longleftarrow$ Trained RF
$SGDC_t \longleftarrow$ Trained SGDC
for i in Corpus **do**
  $C0_{LR} \longleftarrow Pob_{LR} \longleftarrow LR_t(i)$
  $C0_{RF} \longleftarrow Pob_{RF} \longleftarrow RF_t(i)$
  $C0_{SGDC} \longleftarrow Pob_{SGDC} \longleftarrow SGDC_t(i)$
  $C1_{LR} \longleftarrow Pob_{LR} \longleftarrow LR_t(i)$
  $C1_{RF} \longleftarrow Pob_{RF} \longleftarrow RF_t(i)$
  $C1_{SGDC} \longleftarrow Pob_{SGDC} \longleftarrow SGDC_t(i)$
  $AvgProb_{C0} \longleftarrow \frac{C0_{LR}+C0_{RF}+C0_{SGDC}}{3}$
  $AvgProb_{C1} \longleftarrow \frac{C1_{LR}+C1_{RF}+C1_{SGDC}}{3}$
  $SVEM_{Pred} \longleftarrow argmax\{AvgProb_{C0}, AvgProb_{C1}\}$
end for
$Class\ 0|Class\ 1 \longleftarrow SVEM$ prediction

Algorithm 1 depicts the operation of the proposed SVEM model and elucidates its amalgamation of LR, RF, and SGDC for the detection of suicide ideation. In the algorithm, $C0_{LR}$ represents the LR prediction probability for class 0, and the same convention applies to the other models presented. Furthermore, $C1_{LR}$ corresponds to the LR prediction probability for class 1.

3.7 Performance evaluation metrics

When evaluating the performance of an ML model, selecting appropriate evaluation metrics is crucial [58]. Four fundamental outcomes are considered: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These outcomes represent different predictions made by the model. These metrics are then utilized to calculate several commonly used evaluation metrics.

The aggregate of the model's correct forecasts is called its accuracy and is calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision indicates the model's capability to correctly detect positive instances among the predicted positive instances. The formula for precision is

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall (true positive rate or sensitivity) represents the model's ability to correctly identify positive instances among the actual positive instances. Recall is calculated as

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

The F1 score is a combination of recall and precision, a balanced measure approach to assess the model's performance. The F1 score is calculated using the following equation.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

These metrics help assess the model's performance from different perspectives and are commonly used to evaluate classification models.

4 Results and Discussions

Multiple experiments are conducted to assess the effectiveness of both the chosen ML classifiers and the proposed ensemble classifier. The experiments involved the utilization of different feature sets, namely BoW, Glove embeddings, TF-IDF, and TF-IDF and BoW concatenation. The performance of the classifiers was evaluated based on these feature sets.

4.1 Experimental Setup

This section entails the experiment procedure and setup, including the description of the experimental system and the hyperparameters of different classification algorithms. The experiments are conducted on a machine equipped with an Intel Core i7 8th-generation processor, featuring a quad-core setup and 8GB of RAM. The operating system used was Windows 10.0. The implementation was carried out using the Python programming language, utilizing Anaconda 3.0 software and the Jupyter Notebook environment.

4.2 Results of Machine Learning Models

The outcomes presented in this section pertain to ML models' performance using BoW, GloVe, TF-IDF, and the concatenation of BoW and TF-IDF features. The effectiveness of each model is influenced by the specific feature extraction technique employed.

4.2.1 Results Using BoW Features

Table 12 displays the results for all classifiers when using BoW features. The present study employed various ML classifiers with different hyperparameters to achieve higher accuracy. These parameters were selected based on empirical evaluation. For instance, the LR classifier and stochastic gradient descent (SGDC), and RF individually performed well, achieving accuracy scores of 0.92, 0.91, and 0.90, respectively. The proposed ensemble learning model SVEM, outperformed all others, attaining an accuracy score of 0.93. The combination of RF, LR, and SGDC in SVEM harnesses the strengths of these classifiers, promotes diversity, reduces bias and variance, and improves the overall accuracy of the ensemble model. Conversely, the ADA classifier exhibited the lowest accuracy score of 0.88.

Table 12: Results of ML classifiers using BoW features.

Model	Accuracy	Precision	Recall	F1 score
LR	0.92	0.92	0.92	0.92
RF	0.90	0.90	0.90	0.90
SGDC	0.91	0.91	0.91	0.91
NB	0.90	0.91	0.90	0.90
ADA	0.88	0.89	0.88	0.88
SVEM	0.93	0.93	0.93	0.93

4.2.2 Results using TF-IDF Features

Table 13 displays the outcomes of ML models utilizing TF-IDF features. The results indicate a decline in the performance of models when TF-IDF features are employed. Both LR and SGDC achieved an accuracy score of 0.92, and RF achieved an accuracy score of 0.91. Notably, the results of NB and ADA models exhibited degradation when compared to the BoW features. Furthermore, the proposed SVEM achieved an accuracy score of 0.93, indicating that the proposed voting classifier shows a similar performance with both BoW and TF-IDF features.

Table 13: Results of ML classifiers using TF-IDF features.

Model	Accuracy	Precision	Recall	F1 score
LR	0.92	0.92	0.92	0.92
RF	0.91	0.91	0.91	0.91
SGDC	0.92	0.92	0.92	0.92
NB	0.89	0.90	0.89	0.89
ADA	0.87	0.88	0.87	0.87
SVEM	0.93	0.93	0.93	0.93

4.2.3 Results using GloVe Features

Experiments are also carried out using GloVe features with ML models, and results are given in Table 14. Experimental results reveal that the RF model, with an accuracy score of 0.87, performs better with GloVe features, compared to TF-IDF features. Consequently, RF's ability to handle high-dimensional data, its flexibility in capturing complex relationships, and its ensemble approach make it well-suited for leveraging the informative nature of GloVe features and achieving good performance. The accuracy score for LR is 0.83 and SGDC is 0.83, while the proposed SVEM model achieved an accuracy score of 0.85. ADA, with an accuracy score of 0.78, is the least performer on this front.

Table 14: Results of ML classifiers using GloVe features.

Model	Accuracy	Precision	Recall	F1 Score
LR	0.83	0.84	0.83	0.83
RF	0.87	0.87	0.87	0.87
SGDC	0.83	0.84	0.83	0.83
NB	0.79	0.80	0.79	0.79
ADA	0.78	0.79	0.78	0.78
SVEM	0.85	0.86	0.85	0.85

4.2.4 Experimental Results Using Concatenated BoW and TF-IDF Features

By concatenating BoW and TF-IDF features, the objective is to enhance the effectiveness of the feature set, thereby improving the performance of ML models. The results, as shown in Table 15, indicate an overall improvement in the performance of models when used with combined feature sets. In particular, the proposed SVEM model achieves an accuracy of 0.94 when trained on the combined feature set, indicating a significant improvement. Additionally, NB and ADA also outperform previous results, demonstrating the efficacy of the concatenated features. In addition, the performance of RF, LR, and SGDC is enhanced as well, further validating the effectiveness of this feature combination approach.

Table 15: Results of ML classifiers using a hybrid feature set.

Model	Accuracy	Precision	Recall	F1 score
LR	0.92	0.93	0.92	0.92
RF	0.90	0.90	0.90	0.90
SGDC	0.93	0.93	0.93	0.93
NB	0.90	0.91	0.90	0.90
ADA	0.89	0.89	0.89	0.89
SVEM	0.94	0.94	0.94	0.94

Figure 6 shows the comparison between all models' performance using each feature extraction technique in terms of accuracy, precision, recall, and F1 score.

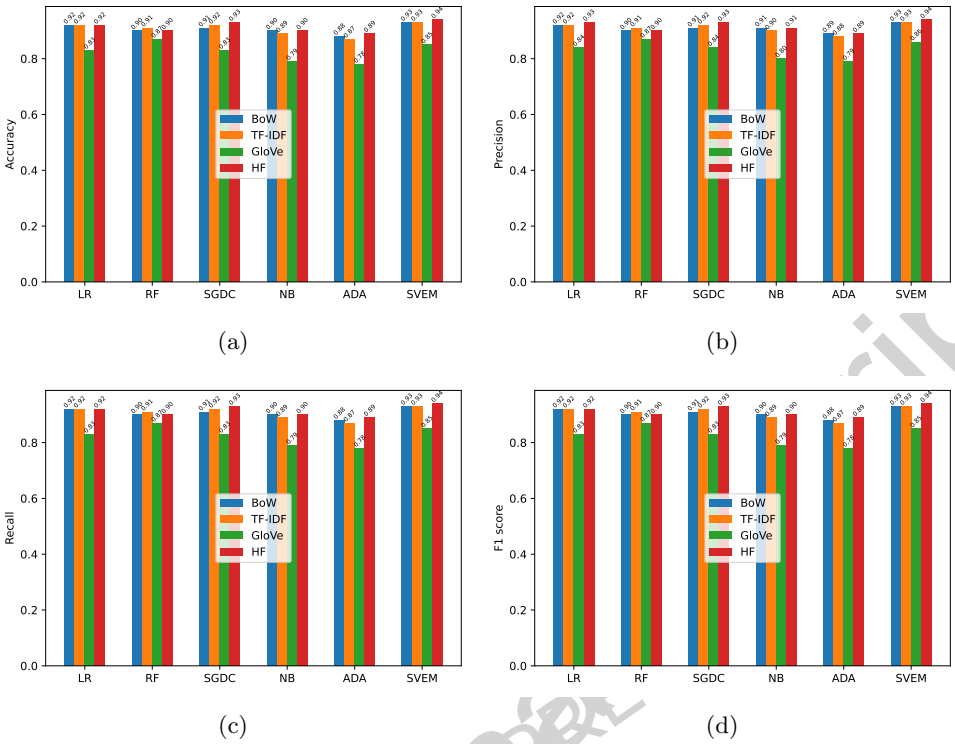

Fig. 6: Performance comparison of models using each feature extraction technique. (**a**) Accuracy; (**b**) Precision; (**c**) Recall and (**d**) F1 score.

4.3 Results of Deep Learning Models

In this study, we deploy four DL models, LSTM, CNN, RNN, and GRU, in comparison with ML models. Table 16 shows the architectures of these DL models. LSTM architecture consists of sequential layers: an embedding layer for converting input data into a numeric representation with 5000 vocabulary size and 100 output dimensions, followed by a dropout layer to prevent overfitting with a 0.5 dropout rate, an LSTM layer with 128 units for capturing temporal dependencies [59], another dropout layer, and finally, a dense layer with 2 units and a softmax activation for classification because we are working for binary classification. Similarly, CNN also uses an embedding layer with the same specifications. After that, a dropout layer to prevent overfitting with a 0.5 dropout rate [60], a 1D convolutional layer (Conv1D) with 128 filters, a ReLU activation function to extract features, and a MaxPooling1D layer for downsampling with

a 4 by 4 pool size is used. A flattened layer to reshape the output into 1 dimension is used, and a dense layer with 2 units and softmax activation for classification [61].

Table 16: Architecture of deep learning models.

LSTM	**CNN**
Embedding (5000, 100) Dropout (0.5) LSTM (128) Dropout (0.5) Dense (2, activation='softmax')	Embedding (5000, 100) Dropout (0.5) Conv1D (128, 4, activation='relu') MaxPooling1D (pool_size=4) Flatten () Dense (2, activation='softmax')
GRU	**RNN**
Embedding (5000, 100) Dropout (0.5) GRU (128) Dropout (0.5) Dense (2, activation='softmax')	Embedding (5000, 100) Dropout (0.5) SimpleRNN (128) Dropout (0.5) Dense (2, activation='softmax')
loss='binary_crossentropy', optimizer='adam', epochs =100	

The GRU model shares similarities with the LSTM design. It includes an embedding layer, a dropout layer, a GRU layer with 128 units, another dropout layer, and a dense layer with 2 units and softmax activation. While RNN also involves an embedding layer, a dropout layer followed by a SimpleRNN layer with 128 units, another dropout layer, and a dense layer with 2 units and softmax activation. All the models are trained with a binary cross-entropy loss function and optimized using the Adam optimizer over 100 epochs. These architectures are designed to process sequential data, each utilizing different recurrent or convolutional structures to learn patterns and features from the input data for suicide prediction tasks.

Table 17 shows the results of DL models. Specifically, LSTM, RNN, and GRU achieved an accuracy score of 0.92, while CNN achieved a slightly lower accuracy score of 0.91. In contrast, the proposed ML-based approach, which utilizes the feature fusion technique, achieved an even higher accuracy score of 0.94. This suggests that the proposed approach, SVEM, outperforms the DL models in terms of accuracy when applied to the given task.

Table 17: Experimental results of DL models.

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Training time (sec)**
GRU	0.92	0.92	0.92	0.92	217.36
CNN	0.91	0.91	0.91	0.91	63.8
LSTM	0.92	0.92	0.92	0.92	235.26
RNN	0.92	0.92	0.92	0.92	223.18

Concerning training time, the CNN model has a lower time due to parallel processing. In comparison, GRU and RNN are slower and balance training time and accuracy. LSTM is the slowest with the highest training time of 235.26 seconds due to its complex architecture.

4.4 Performance Comparison of Models on Suicide Ideation Dataset

The publicly available Suicide Ideation Dataset is used for our models to evaluate their performance. A performance comparison between ML and DL models is carried out in Table 18. Results show that the proposed model has performed better than other models. The proposed SVEM model shows better performance for suicide ideation detection compared to other ML and DL models employed in this study, thereby indicating the superiority of the proposed model.

Table 18: Performance comparison of ML and DL models used in this study.

Machine learning					**Deep Learning**	
Model	**Accuracy**				**Model**	**Accuracy**
	BoW	TF-IDF	Glove	HF		
LR	0.92	0.92	0.83	0.92	GRU	0.92
RF	0.90	0.91	0.87	0.90	CNN	0.91
SGDC	0.91	0.92	0.83	0.93	LSTM	0.92
NB	0.90	0.89	0.79	0.90	RNN	0.92
ADA	0.88	0.87	0.78	0.89		
SVEM	0.93	0.93	0.85	0.94		

Figure 7 shows confusion matrix results representing the performance evaluation of the best model with each feature extraction technique. In each matrix, the diagonal values correspond to correctly classified instances, while the off-diagonal values represent misclassification. For the SVEM model, 32922 were accurately classified as non-suicide out of a total of 69623 instances, and 32232 instances were correctly classified as suicide. So in total, SVEM gives 65154 correct predictions. There were 1950 false negatives (misclassified instances that are non-suicide class samples), and 2519 false positives (misclassified instances that are suicide class samples). So in total, it gives 4469 wrong predictions. In the case of BoW features, SVEM performs well, as the model accurately classified 33075 non-suicide instances and 31297 suicide instances, while misclassifying 1797 instances as false negatives and 3454 instances as false positives. On the other hand, RF performs well with the GloVe feature as compared to SVEM. RF model gives 60526 correct predictions and 9097 wrong predictions. Lastly, the SVEM model performs well with TF-IDF features; it gives 64630 correct predictions, along with 4993 wrong predictions.

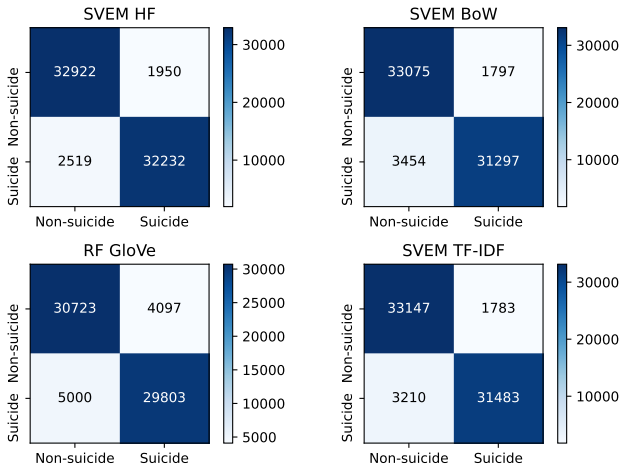

Fig. 7: Confusion matrices for the best models with each feature set.

Figure 8 shows the number of correct and wrong predictions for the best models with each feature extraction technique. SVEM with hybrid features shows the highest number of correct predictions and the lowest number of wrong predictions.

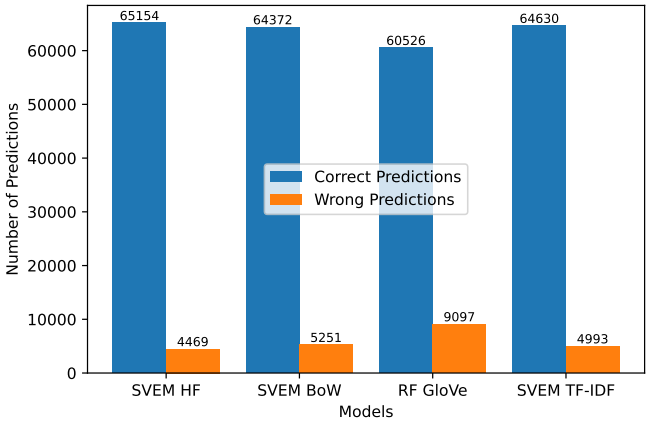

Fig. 8: Correct and wrong predictions by the best models with each feature set.

Table 19 shows per-class performance metrics for the best-performing model with each feature extraction technique. Each model's precision, recall, and F1 score values are reported for both the 'non-suicide' and 'suicide' classes. The SVEM with feature fusion, SVEM with BoW, and SVEM with TF-IDF demonstrated consistent performance across classes, achieving precision and recall scores of around 0.93 to 0.95, resulting in balanced F1 scores of approximately 0.93 to 0.94. In contrast, the RF

GloVe model exhibited slightly lower precision and recall values, particularly for the "non-suicide" class, resulting in an overall F1 score of about 0.87. These per-class results also show the significance of the proposed SVEM model.

Table 19: Per-class performance of the proposed model.

Model	Class	Precision (±SD)	Recall (±SD)	F1 score (±SD)
SVEM HF	non-suicide	0.93 (±0.02)	0.94 (±0.02)	0.94 (±0.02)
	suicide	0.94 (±0.01)	0.93 (±0.02)	0.94 (±0.02)
SVEM BoW	non-suicide	0.93 (±0.02)	0.94 (±0.01)	0.94 (±0.02)
	suicide	0.94 (±0.02)	0.93 (±0.02)	0.94 (±0.01)
SVEM TF-IDF	non-suicide	0.91 (±0.03)	0.95 (±0.01)	0.93 (±0.02)
	suicide	0.95 (±0.01)	0.91 (±0.03)	0.93 (±0.02)
RF GloVe	non-suicide	0.86 (±0.04)	0.88 (±0.03)	0.87 (±0.03)
	suicide	0.88 (±0.03)	0.86 (±0.03)	0.87 (±0.02)

4.5 TF-IDF+BoW vs GloVe

The GloVe approach learns vectors from global co-occurrences of words, which indicates that the words used in a similar context have the same representations. Since it is context-dependent, the same vector is used for a word, even when it is used in a different context. Due to this limitation, it might miss important cues, particularly when used in text containing emotions. Similarly, it shows less sensitivity to those words that are important but rare, for example, those used to indicate mental health or condition. In the current study, since the data used contains emotion and mental states, it shows poor performance compared to the ensemble of BoW and TF-IDF.

Pre-trained DL embeddings may not necessarily capture all of the domain-specific fine-grained nuances. They need a larger corpus for better generalization. In addition, the risk of overfitting is high when smaller or domain-specific datasets are used. TF-IDF and BoW can perform better than Word2Vec, GloV, etc., for different domain texts. TF-IDF and BoW do not need large datasets to learn good representations. Although both BoW and TF-IDF are task-specific and data-driven, they learn from the actual word distribution in the data, which tends to perform better in low-resource environments.

4.6 Results of K-fold Cross-validation

To demonstrate the significance of the proposed approach, the k-fold cross-validation is performed. K-fold cross-validation is a technique used to evaluate how well an ML model will perform on unseen data [62]. Instead of just splitting your dataset once into training and testing sets, it splits the data into K equal parts (folds) and tests the model K times, each time using a different fold as the test set and the remaining folds for training [63].

The consistent performance of SVEM across each fold underscores the robustness of the model throughout the dataset. We employed a 10-fold cross-validation and assessed the mean accuracy and standard deviation (SD), as presented in Table 20. SVEM exhibits noteworthy performance with HF features, achieving an average accuracy of 0.93 with an SD of 0.01.

Table 20: 10-fold cross-validation using SVEM.

Fold	BoW	TF-IDF	GloVe	HF
1	0.91	0.91	0.84	0.90
2	0.90	0.92	0.83	0.93
3	0.90	0.91	0.81	0.94
4	0.89	0.90	0.85	0.91
5	0.89	0.92	0.78	0.94
6	0.88	0.87	0.81	0.93
7	0.91	0.93	0.84	0.94
8	0.92	0.92	0.83	0.95
9	0.91	0.93	0.85	0.94
10	0.93	0.93	0.86	0.94
Mean	0.90	0.91	0.83	0.93
SD	0.01	0.01	0.02	0.01

4.7 Testing Generalization of Proposed Approach Using an Imbalanced Dataset

To demonstrate the significance of the proposed approach in terms of generalization, we conducted experiments on an imbalanced dataset. We deliberately made our dataset imbalanced, selecting 50,000 'suicide' samples and 15,000 'non-suicide' samples. Subsequently, we conducted experiments using our proposed approach. This intentionally imbalanced dataset provides an appropriate environment to test the generalization of models. The proposed approach demonstrates significant results even with an imbalanced dataset, achieving an accuracy of 0.94 and a notable F1 score of 0.91, surpassing other approaches as shown in Table 21.

Table 21: Proposed approach results using an imbalanced dataset.

Features	Accuracy	Class	Precision	Recall	F1 Score
TF-IDF	0.93	non-suicide	0.83	0.91	0.86
		suicide	0.95	0.93	0.94
		Avg.	0.89	0.92	0.90
BoW	0.93	non-suicide	0.81	0.90	0.85
		suicide	0.97	0.94	0.95
		Avg.	0.89	0.92	0.90
GloVE	0.84	non-suicide	0.64	0.79	0.71
		suicide	0.93	0.86	0.90
		Avg.	0.79	0.83	0.80
HF	0.94	non-suicide	0.91	0.81	0.86
		suicide	0.95	0.98	0.96
		Avg.	0.93	0.92	0.91

4.8 Comparison With Other Ensemble Approaches

In this section, we conducted a comparison between our proposed approach and other ensemble methodologies. Firstly, we implemented the Hard Voting Ensemble Model (HVEM) utilizing LR, RF, and SGDC models [64]. Additionally, we utilized the Stack Ensemble Model (SEM) by employing LR and SGDC as base learners and RF as the meta-learner [65]. Furthermore, we created an ensemble comprising LSTM-GRU-RNN models and combined them sequentially in the specified order [66]. We deploy other proposed approaches using HF. Table 22 presents a comparison of ensemble methods, indicating that HVEM performs very similarly to SVEM in terms of accuracy but exhibits slightly lower precision compared to SVEM. HVEM generates 22 more incorrect predictions than SVEM. However, SEM and LSTM-GRU-RNN both display poorer performance in comparison to SVEM and HVEM.

SVEM stands out significantly compared to HVEM and SEM. Soft voting, employed in SVEM, involves combining the predicted probabilities from different models and averaging them to make a final prediction. This method is often considered more robust than hard voting because it considers the confidence levels or probabilities assigned by each model to its predictions.

Table 22: Experimental results using different ensemble approaches.

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**
HVEM	0.94	0.93	0.94	0.94
SEM	0.93	0.93	0.93	0.93
LSTM-GRU-RNN	0.92	0.92	0.92	0.92
SVEM	0.94	0.94	0.94	0.94

4.9 Results using Pre-trained Embedding BERT

We also compare results with pre-trained embedding in comparison with our HF approach. Using BERT embeddings (https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2), the performance of all ML models improved notably due to BERT's deep contextual understanding of language, as shown in Table 23. Among the models evaluated, LR and SGD Classifier achieved the highest accuracy (0.9325), with strong macro precision, recall, and F1-scores (all 0.93), demonstrating their effectiveness in capturing the semantic nuances encoded by BERT. The SVEM ensemble model, which combines LR, RF, and SGD, also performed comparably with an accuracy of 0.9310, confirming the benefit of combining complementary classifiers. RF and ADA followed closely with accuracies of 0.9100 and 0.9120, respectively, showing that ensemble-based methods can leverage BERT's embeddings effectively, though slightly less consistently. NB, however, lagged with an accuracy of 0.8785, as it is less suited for dense, continuous embeddings like those produced by BERT. Overall, BERT significantly enhances the classification of suicide-related content by providing richer, context-aware representations, and models like LR, SGDC, and SVEM make the most of this capability.

Table 23: Performance comparison using BERT embeddings.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9325	0.93	0.93	0.93
Random Forest	0.9100	0.91	0.91	0.91
SGD Classifier	0.9325	0.93	0.93	0.93
Naive Bayes	0.8785	0.88	0.88	0.88
AdaBoost	0.9120	0.91	0.91	0.91
SVEM Ensemble	0.9310	0.93	0.93	0.93

4.10 LIME Results

The use of XAI is important for understanding the rationale of decisions made by the ML model, which in turn increases transparency [67]. In this regard, LIME, a well-known approach, is commonly used [68]. Figure 9 shows LIME's explanation that the model predicted the input text as "suicide" with confidence of 72%, compared to 28% for "non-suicide." The highlighted words in the text indicate which terms contributed most to the classification. Words such as "kill", "tried", "know", and "dont" are shaded in orange, meaning they strongly influenced the model towards the suicide class, with "kill" contributing the most (0.20). Conversely, words like "english", "phone", and "information" are shaded in blue, contributing slightly toward the non-suicide classification, though with lower influence. This local explanation demonstrates how emotionally and contextually intense words drive the model's decision-making, helping to interpret its reasoning in a transparent and human-understandable way.

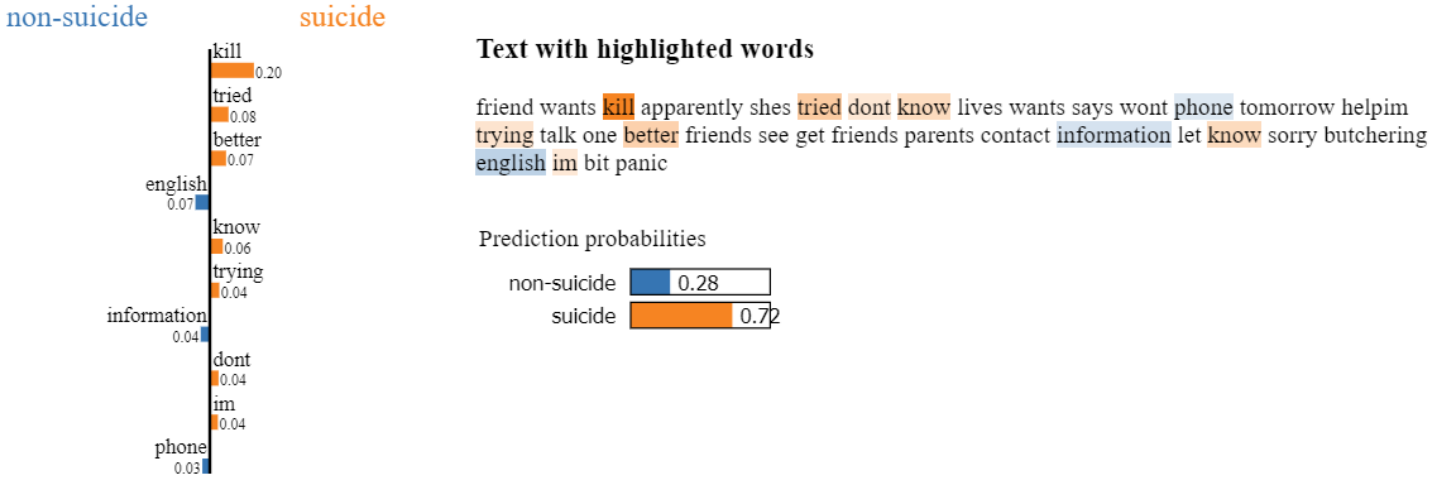

Fig. 9: Results using LIME XAI model.

4.11 Statistical Test

In this section, we discussed the statistical significance of the proposed study.

To evaluate the statistical significance of SVEM's performance compared to other ML models, we conducted an independent two-sample Student's T-test. We collected the accuracy values of SVEM across four feature extraction techniques (BoW, TF-IDF, GloVe, and HF) and compared them with the combined accuracy scores of five other models (LR, RF, SGDC, NB, and ADA) under the same conditions. The T-test was performed under the assumption of unequal variances to account for possible distributional differences. The statistical analysis highlights that SVEM consistently achieves higher accuracy across all feature extraction methods compared to traditional ML models. Although the Student's T-test yielded a p-value of 0.265 as shown in Table 24, indicating the performance difference is not statistically significant at standard thresholds (α=0.01 to 0.10), SVEM's consistent outperformance suggests a practical advantage. When interpreted at a more lenient threshold (α=0.50), the improvement becomes statistically significant, reflecting SVEM's potential as a more robust and reliable model.

Table 24: Interpretation of SVEM's T-test results at varying significance levels.

Significance Level (α)	P-value	Interpretation
0.01	0.265	Fail to Reject H_0
0.05	0.265	Fail to Reject H_0
0.10	0.265	Fail to Reject H_0
0.20	0.265	Fail to Reject H_0
0.25	0.265	Fail to Reject H_0
0.30	0.265	Reject H_0
0.40	0.265	Reject H_0
0.50	0.265	Reject H_0
0.60	0.265	Reject H_0
0.70	0.265	Reject H_0

4.12 Big-O Complexity

To assess the computational efficiency of SVEM, we analyze the Big-O time complexity based on the feature representations and dataset characteristics. SVEM operates on classical text features such as TF-IDF, BoW, and GloVe, and its total training time complexity is approximated as $O(n \cdot d \cdot i + n \cdot d + t \cdot m \cdot n \cdot \log n)$. Here, n represents the number of training instances (e.g., 232,074 in this case), d is the number of features (terms generated by vectorizers), i is the number of iterations for SGDC, t is the number of trees in the Random Forest (e.g., 300), and m is the number of features considered per split (often $\sqrt{d}$ or a fraction of d). The SGDC and LR components scale linearly with the number of features and samples, while the RF component adds additional overhead due to its ensemble of deep trees.

In contrast, DL models such as RNN, LSTM, GRU, and CNN require significantly more computational power. Recurrent architectures have a training complexity of

$O(n \cdot s \cdot h^2)$, where s is the average sequence length (number of tokens per input text) and h is the size of the hidden layer (typically between 128 and 512). CNNs for text classification operate with a complexity of $O(n \cdot f \cdot k \cdot d)$, where f is the number of filters, k is the kernel size, and d is the embedding dimension. These models also require GPU acceleration for efficient training, particularly when working with large datasets or long input sequences.

4.13 Performance Comparison with Existing Studies

The performance of the proposed approach is compared with previous techniques to demonstrate its significance. Table 25 provides a comparison between the SVEM model and recent state-of-the-art approaches. In this regard, recent studies are selected from the existing literature. In particular, those recent studies are selected that utilized the same benchmark dataset that is used in this study. Many of these studies employed DL-based architectures, which are computationally expensive. However, despite high computational complexity, the results are not better than the current study. For instance, in the study by Renjith et al. [28], an LSTM-attention-CNN model was used for suicide ideation detection, yielding an accuracy score of 90.3%.

Similarly, the SVM model for suicide detection in [28] provides an 87% accuracy. Both [30] and [29] utilized DL models and obtained 91.5% and 90.8% accuracy, respectively. To the best of our knowledge, no previous study on suicide ideation detection employed a hybrid feature engineering approach for model training.

Another recent study is [69], which utilized a number of ML and BERT models for suicide detection. The authors used RF, Bernoulli NB, LR, RoBERTA, DeBERTa, DistBERT, and SquuezeBERT architectures. The highest accuracy of 93.5% is reported using LR. Another similar work using the same dataset is from et al. [70], which used NB, SVM, DT, LR, RF, and XGBoost, with the best performance reported by the NB model. Using an ensemble learning approach at the learning level and hybrid features for model training yields the best results with a 94% accuracy in this study.

Table 25: Performance comparison of the proposed approach with state-of-the-art models.

Ref.	Year	Model	Accuracy
[28]	2022	LSTM-attention-CNN	90.3%
[30]	2022	SVM	87%
[71]	2022	CNN-Bi-LSTM	91.5%
[29]	2023	LSTM	90.8%
[69]	2024	RF, Bernoulli NB, LR, RoBERTA, DeBERTa, DistBERT and SquuezeBERT	93.50% using LR
[70]	2024	NB, SVM 0.85, DT, LR, RF, XGBoost	85.00% using SVM
Current Study	2025	SVEM	94.10%

4.14 Discussions

This study presents an ensemble approach combining the strengths of LR and RF as base models, while the SGDC is used as a meta classifier. In addition, a hybrid feature engineering approach is also designed, utilizing TF-IDF and BoW for better training of the models. Despite being complicated, BERT-based embeddings did not surpass the hybrid TF-IDF + BoW ensemble. This is largely due to the noisy, short, and domain-specific nature of social media posts, which reduces the advantage of deep contextual embeddings. In contrast, hybrid handcrafted features capture frequency and weighted term distributions more effectively for such sparse text, enabling the ensemble to achieve higher stability, interpretability, and lower computational cost.

The proposed SVEM showed better performance compared to DL models like RNN, GRU, CNN, etc. Suicide ideation datasets, such as those from Reddit and Twitter, are usually small. DL models need a large-scale annotated dataset for better generalization. On the other hand, LR, RF, and SGDC perform better on smaller datasets. In addition, combining LR and RF as base learners leads to a mix of linear decision boundaries and non-linear decision rules, leading to capturing both local and global trends in text data.

DL models are not good with sparse features, found in textual data such as from social media platforms. Models like LR and SGDC are good at handling sparse, high-dimensional input without overfitting. A balanced dataset, used in this study, also helps in this regard. On the other hand, DL models often overfit on smaller datasets, particularly those involving subtle linguistic cues such as those containing sarcastic or suicidal emotions.

4.15 Real-World Applications

Suicide ideation detection systems are very important, particularly in the current era of digitization and social media platforms. The following are a few real-world applications of such systems.

- Such systems can be used in healthcare and mental state monitoring systems, particularly in clinical screening tools used by psychiatrists. Similarly, systems involving electronic health record (EHR) systems involving ML and NLP can use such systems for warning signs.
- Suicide ideation detection systems can be used to monitor social media for analyzing text on social media for possible detection of posts indicating suicide threats.
- Education puts great pressure on students and affects their mental state. Such systems can be incorporated into the university education system for students' mental well-being.
- Suicide ideation detection can also be utilized by the public sector to monitor the overall mental state of the public and design effective policies to mitigate suicide risks.

4.16 Implications of Proposed Approach

The proposed approach for suicide ideation detection has several implications.

- Early and timely identification of individuals at risk can help psychologists, psychiatrists, and support organizations to identify those at risk of suicide and intervene before self-harm occurs.
- It can help healthcare systems prioritize cases requiring urgent attention and can improve efficiency in suicide prevention strategies.
- The proposed approach can be used to provide data-driven understanding of suicide trends, risk factors, and vulnerable groups.
- SVEM can be adopted for real-time monitoring of students under an educational burden. It can be used for the development of scalable and automated systems for continuous screening of social media platforms.

4.17 Limitations

DL models are limited due to several factors. The limitation of poor results and lack of generalization is due to various factors, such as the lack of pre-trained contextual embeddings, poor hyperparameter tuning, and social media language's inherent noise, which can undermine the performance of DL models that are context-dependent. The study also acknowledges limitations in terms of dataset bias, with the data being mainly acquired from one platform (Reddit), which may compromise generalizability to various user groups and social media settings. Model interpretability is also an issue, particularly with DL models, which are typically black boxes and problematic in applications in mental health, where transparency is essential.

In addition, demographic bias can arise due to the overrepresentation of specific groups with respect to ethnicity, language, or platform, and minority groups can be underrepresented. Generalizability problems can also be part of the DL model due to the utilization of data from a single platform. For example, a model trained using data from X (Twitter) might not show the same efficacy when tested on Reddit data and vice versa. Future work will involve leveraging data from multiple platforms, using pre-trained language models, and making use of more XAI techniques to enhance model transparency.

4.18 Ethical Concerns

Mining sensitive social media data raises concerns of consent, surveillance, and data misuse. In addition, handling ethical concerns, often by false positives, is a sensitive issue and a grave concern, particularly in systems involving the assessment of mental health. A false positive in this case might lead to distress, depression, and the victim's mistrust of the person falsely flagged as having suicidal thoughts. In addition, in the case of involving third-party intervention, it can lead to privacy breaches. One widely used solution is the human-in-the-loop approach, where such flags are reviewed by a human medical expert before any response. The medical expert gives clinical judgments. In addition, multi-stage filtering systems, which involve multiple layers, are utilized. Another potential solution is the development of systems that balance sensitivity and specificity. Regular retraining, although computationally overhead, can ensure this balance. Getting feedback from the user is also used in several mental health monitoring systems.

The use of automated suicide ideation detection has life-saving potential, but it is also accompanied by several challenges and concerns. Therefore, its deployment must strike a balance between technological innovation and ethical safeguards.

5 Conclusions

This study proposed an ensemble approach and conducted experiments for detecting suicidal ideation using social media posts from Reddit. The ensemble model was employed with combined features from TF-IDF and BoW approaches. Trained on richer feature sets, the ensemble architecture substantially enhanced performance, achieving an accuracy of 94%. Compared to deep learning and pre-trained models such as BERT, the ensemble method delivered significantly better accuracy while maintaining low computational costs. By leveraging TF-IDF and BoW, the model captured both weighted and frequency-based features, leading to improved performance over GloVe. The use of the LIME explainable AI approach further highlighted the importance of different features for suicidal ideation detection. Overall, the study demonstrated that hybrid features provide richer textual representations than individual methods, thereby improving model performance. Ensemble learning also enables the capture of more complex patterns due to the diversity of base learners, such as the combination of linear and tree-based models. Furthermore, explainability is a critical factor in complex and sensitive applications like suicidal ideation detection, and LIME offers valuable insights into decision-making, helping evaluators better justify model outcomes. The findings highlight the practical potential of combining ensemble learning with explainable AI for early detection of suicidal ideation. Such a framework could be applied in real-world mental health monitoring systems, providing healthcare professionals with supportive tools for timely intervention.

The strengths of this research lie in developing a lightweight yet high-performing ensemble model, the effective use of hybrid features, and the incorporation of explainability, which is essential in sensitive domains such as suicide ideation. However, certain limitations remain. The dataset is restricted to Reddit posts, which may introduce bias and limit generalizability across platforms and populations. Moreover, while the ensemble approach enhances performance, its reliance on handcrafted features may reduce scalability compared to advanced deep learning models. In future work, we plan to expand the dataset by incorporating posts from multiple social media platforms to improve generalization, explore larger pre-trained BERT models, and further strengthen the robustness of our approach.

Declarations

Funding

This research is funded by the European University of Atlantic.

Conflict of Interest

"The authors declare no conflict of interests."

Ethics approval

"Not applicable."

Consent to participate

"Not applicable."

Consent for publication

"Not applicable."

Availability of data and materials

"The datasets used and/or analysed during the current study available from the corresponding author on reasonable request."

Code availability

"Not applicable."

Authors' contributions

E.K. conceived the idea, performed data curation and wrote the original manuscript.
J.G.C. designed methodology, performed formal analysis and validation.
A.I. performed formal analysis and data curation, and conceived the idea.
R.H. designed methodology, dealt with software and performed visualization.
M.G.V. dealt with software, performed visualization and investigation.
E.S.A. performed investigation and visualization, and acquired funding.
I.d.l.T.D. performed validation, provided resources and carried out formal analysis.
I.A. supervised the work, performed validation and edited the manuscript.
All authors reviewed the manuscript and approved it.

References

[1] WHO, R.: WHO. Suicide.https://www.who.int/health-topics/suicide#tab=tab_1 [Accessed: (22july2023)] (2022)

[2] Bilsen, J.: Suicide and youth: risk factors. Frontiers in psychiatry **9**, 540 (2018)

[3] Värnik, P.: Suicide in the world. International journal of environmental research and public health **9**(3), 760–771 (2012)

[4] Hawton, K., Van Heeringen, K.: The International Handbook of Suicide and Attempted Suicide. John Wiley & Sons, ??? (2000)

[5] World Health Organization: Preventing suicide: A global imperative. https://www.who.int/publications/i/item/9789241564779 (2014)

[6] O'Connor, R.C., Nock, M.K.: The psychology of suicidal behaviour. The Lancet Psychiatry **1**(1), 73–85 (2014)

[7] American Foundation for Suicide Prevention: Risk Factors, Protective Factors, and Warning Signs. American Foundation for Suicide Prevention. https://afsp.org/risk-factors-protective-factors-and-warning-signs/ note=Accessed on 23 April, 2025 (2022)

[8] Franklin, J.C., Ribeiro, J.D., Fox, K.R., Bentley, K.H., Kleiman, E.M., Huang, X., Musacchio, K.M., Jaroszewski, A.C., Chang, B.P., Nock, M.K.: Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychological bulletin **143**(2), 187 (2017)

[9] Castillo-Sánchez, G., Marques, G., Dorronzoro, E., Rivera-Romero, O., Franco-Martín, M., Torre-Díez, I.: Suicide risk assessment using machine learning and social networks: a scoping review. Journal of medical systems **44**(12), 205 (2020)

[10] Aladağ, A.E., Muderrisoglu, S., Akbas, N.B., Zahmacioglu, O., Bingol, H.O.: Detecting suicidal ideation on forums: proof-of-concept study. Journal of medical Internet research **20**(6), 9840 (2018)

[11] Harmer, B., Lee, S., Rizvi, A., Saadabadi, A.: Suicidal Ideation. StatPearls Publishing, Treasure Island (FL), ??? (2025). http://europepmc.org/books/NBK565877

[12] Simon, R.I.: Passive suicidal ideation: Still a high-risk clinical scenario. Current Psychiatry **13**(3), 13–15 (2014)

[13] Weber, A.N., Michail, M., Thompson, A., Fiedorowicz, J.G.: Psychiatric emergencies: assessing and managing suicidal ideation. Medical Clinics **101**(3), 553–571 (2017)

[14] Gaur, M., Aribandi, V., Alambo, A., Kursuncu, U., Thirunarayan, K., Beich, J., Pathak, J., Sheth, A.: Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs. PloS one **16**(5), 0250448 (2021)

[15] Twenge, J.M.: Increases in depression, self-harm, and suicide among us adolescents after 2012 and links to technology use: possible mechanisms. Psychiatric Research and Clinical Practice **2**(1), 19–25 (2020)

[16] Green, T., Wilhelmsen, T., Wilmots, E., Dodd, B., Quinn, S.: Social anxiety, attributes of online communication and self-disclosure across private and public facebook communication. Computers in Human Behavior **58**, 206–213 (2016)

[17] Abboute, A., Boudjeriou, Y., Entringer, G., Azé, J., Bringay, S., Poncelet, P.: Mining twitter for suicide prevention. In: Natural Language Processing and Information Systems: 19th International Conference on Applications of Natural

Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings 19, pp. 250–253 (2014). Springer

[18] Chiang, W.-C., Cheng, P.-H., Su, M.-J., Chen, H.-S., Wu, S.-W., Lin, J.-K.: Socio-health with personal mental health records: suicidal-tendency observation system on facebook for taiwanese adolescents and young adults. In: 2011 IEEE 13th International Conference on e-Health Networking, Applications and Services, pp. 46–51 (2011). IEEE

[19] Amin, S.U., Jung, Y., Fayaz, M., Kim, B., Seo, S.: Enhancing pine wilt disease detection with synthetic data and external attention-based transformers. Engineering Applications of Artificial Intelligence **159**, 111655 (2025)

[20] Amin, S.U., Abbas, M.S., Kim, B., Jung, Y., Seo, S.: Enhanced anomaly detection in pandemic surveillance videos: An attention approach with efficientnet-b0 and cbam integration. IEEE Access (2024)

[21] Ul Amin, S., Kim, Y., Sami, I., Park, S., Seo, S.: An efficient attention-based strategy for anomaly detection in surveillance video. Computer Systems Science & Engineering **46**(3) (2023)

[22] Ul Amin, S., Kim, B., Jung, Y., Seo, S., Park, S.: Video anomaly detection utilizing efficient spatiotemporal feature fusion with 3d convolutions and long short-term memory modules. Advanced Intelligent Systems **6**(7), 2300706 (2024)

[23] Abdelmalak, M.E.S., Gaber, K.S., Ahmed, M.A., OubeBlika, N., Zaki, A.M., Eid, M.M.: Ber-xgboost: pothole detection based on feature extraction and optimized xgboost using ber metaheuristic algorithm. J Artif Intell Metaheuristics **6**(2), 46–55 (2023)

[24] Jabbar, A., Liaqat, H.B., Akram, A., Sana, M.U., Azpíroz, I.D., Diez, I.D.L.T., Ashraf, I.: A lesion-based diabetic retinopathy detection through hybrid deep learning model. IEEE Access (2024)

[25] Ashraf, I., Hur, S., Park, Y.: Indoor positioning on disparate commercial smartphones using wi-fi access points coverage area. Sensors **19**(19), 4351 (2019)

[26] Mujahid, M., Kanwal, K., Rustam, F., Aljedaani, W., Ashraf, I.: Arabic chatgpt tweets classification using roberta and bert ensemble model. ACM Transactions on Asian and Low-Resource Language Information Processing **22**(8), 1–23 (2023)

[27] ZainEldin, H., Gamel, S.A., El-Kenawy, E.-S.M., Alharbi, A.H., Khafaga, D.S., Ibrahim, A., Talaat, F.M.: Brain tumor detection and classification using deep learning and sine-cosine fitness grey wolf optimization. Bioengineering **10**(1), 18 (2022)

[28] 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. Journal of King Saud University-Computer and Information Sciences **34**(10), 9564–9575 (2022)

[29] Deepa, J., Shriraaman, S., Shruti, V., Vasanth, G.: Detecting and determining degree of suicidal ideation on tweets using lstm and machine learning models. Journal of Survey in Fisheries Sciences **10**(2S), 3217–3224 (2023)

[30] Chatterjee, M., Kumar, P., Samanta, P., Sarkar, D.: Suicide ideation detection from online social media: A multi-modal feature based technique. International Journal of Information Management Data Insights **2**(2), 100103 (2022)

[31] Agarwal, .D.B. K: Deep learning based approach for detecting suicidal ideation in hindi-english code-mixed text: Baseline and corpus. In Proceedings of the 18th International Conference on Natural Language Processing (ICON), 100–105 (2021)

[32] Bandari, N., Kancharla, M., Kunsoth, U.: Suicidal tweets detection in online social media using machine learning. Turkish Journal of Computer and Mathematics Education (TURCOMAT) **13**(03), 1258–1267 (2022)

[33] Ji, S., Yu, C.P., Fung, S.-f., Pan, S., Long, G.: Supervised learning for suicidal ideation detection in online user content. Complexity **2018**(1), 6157249 (2018)

[34] Nordin, N., Zainol, Z., Mohd Noor, M.H., Lai Fong, C.: A comparative study of machine learning techniques for suicide attempts predictive model. Health informatics journal **27**(1), 1460458221989395 (2021)

[35] Liu, J., Shi, M., Jiang, H.: Detecting suicidal ideation in social media: An ensemble method based on feature fusion. International journal of environmental research and public health **19**(13), 8197 (2022)

[36] Tadesse, M.M., Lin, H., Xu, B., Yang, L.: Detection of suicide ideation in social media forums using deep learning. Algorithms **13**(1), 7 (2019)

[37] Gupta, V., Jain, N., Katariya, P., Kumar, A., Mohan, S., Ahmadian, A., Ferrara, M.: An emotion care model using multimodal textual analysis on covid-19. Chaos, Solitons & Fractals **144**, 110708 (2021)

[38] Sharma, D., Gupta, V., Singh, V.K.: Detection of homophobia & transphobia in malayalam and tamil: Exploring deep learning methods. In: Advanced Network Technologies and Intelligent Computing, pp. 217–226. Springer, Cham (2023)

[39] Arora, S., Agrawal, V., Kumar, D., Arora, S., Banshal, S.K.: Sentimental impact of fake news on social media using an integrated ensemble framework. Social Network Analysis and Mining **14**(1), 185 (2024)

[40] Qorich, M., El Ouazzani, R.: Advanced deep learning and large language models for suicide ideation detection on social media. Progress in Artificial Intelligence **13**(2), 135–147 (2024)

[41] Pourkeyvan, A., Safa, R., Sorourkhah, A.: Harnessing the power of hugging face transformers for predicting mental health disorders in social networks. IEEE Access **12**, 28025–28035 (2024)

[42] Li, A.: Predicting negative attitudes towards suicide in social media texts: prediction model development and validation study. Frontiers in public health **12**, 1401322 (2024)

[43] Kaminsky, Z., McQuaid, R.J., Hellemans, K.G., Patterson, Z.R., Saad, M., Gabrys, R.L., Kendzerska, T., Abizaid, A., Robillard, R.: Machine learning–based suicide risk prediction model for suicidal trajectory on social media following suicidal mentions: Independent algorithm validation. Journal of Medical Internet Research **26**, 49927 (2024)

[44] Komati, N.: Suicide and Depression Detection

[45] Ishaq, A., Umer, M., Mushtaq, M.F., Medaglia, C., Siddiqui, H.U.R., Mehmood, A., Choi, G.S.: Extensive hotel reviews classification using long short term memory. Journal of Ambient Intelligence and Humanized Computing **12**, 9375–9385 (2021)

[46] Rustam, F., Mehmood, A., Ahmad, M., Ullah, S., Khan, D.M., Choi, G.S.: Classification of shopify app user reviews using novel multi text features. IEEE Access **8**, 30234–30244 (2020)

[47] Qader, W.A., Ameen, M.M., Ahmed, B.I.: An overview of bag of words; importance, implementation, applications, and challenges. In: 2019 International Engineering Conference (IEC), pp. 200–204 (2019). IEEE

[48] Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS) **26**(3), 1–37 (2008)

[49] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[50] Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., Choi, G.S.: A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis. Plos one **16**(2), 0245909 (2021)

[51] Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression vol. 398. John Wiley & Sons, ??? (2013)

[52] Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R., Kundu, S.: Improved random forest for classification. IEEE Transactions on Image Processing **27**(8), 4012–4024 (2018)

[53] Gaye, B., Zhang, D., Wulamu, A.: Sentiment classification for employees reviews using regression vector-stochastic gradient descent classifier (rv-sgdc). PeerJ Computer Science **7**, 712 (2021)

[54] Wyner, A.J., Olson, M., Bleich, J., Mease, D.: Explaining the success of adaboost and random forests as interpolating classifiers. The Journal of Machine Learning Research **18**(1), 1558–1590 (2017)

[55] Jahromi, A.H., Taheri, M.: A non-parametric mixture of gaussian naive bayes classifiers based on local independent features. In: 2017 Artificial Intelligence and Signal Processing Conference (AISP), pp. 209–212 (2017). IEEE

[56] Kulkarni, V.Y., Sinha, P.K.: Random forest classifiers: a survey and future research directions. Int. J. Adv. Comput **36**(1), 1144–1153 (2013)

[57] Gladence, L.M., Karthi, M., Anu, V.M.: A statistical comparison of logistic regression and different bayes classification methods for machine learning. ARPN Journal of Engineering and Applied Sciences **10**(14), 5947–5953 (2015)

[58] Lee, E., Rustam, F., Shahzad, H.F., Washington, P.B., Ishaq, A., Ashraf, I.: Drug usage safety from drug reviews with hybrid machine learning approach. Computer Systems Science & Engineering **46**(1) (2023)

[59] Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of cnn and rnn for natural language processing. arXiv preprint arXiv:1702.01923 (2017)

[60] Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H.A., Bisogni, C.: Rfcnn: Traffic accident severity prediction based on decision level fusion of machine and deep learning model. IEEE Access **9**, 128359–128371 (2021)

[61] Fu, R., Zhang, Z., Li, L.: Using lstm and gru neural network methods for traffic flow prediction. In: 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), pp. 324–328 (2016). IEEE

[62] Vasant Bidwe, R., Mishra, S., Kamini Bajaj, S., Kotecha, K.: Attention-focused eye gaze analysis to predict autistic traits using transfer learning. International Journal of Computational Intelligence Systems **17**(1), 120 (2024)

[63] Bidwe, R., Mishra, S., Bajaj, S., Kotecha, K.: Leveraging hybrid model of convnextbase and lightgbm for early asd detection via eye-gaze analysis. MethodsX **14**, 103166 (2025)

[64] Aslam, N., Xia, K., Rustam, F., Lee, E., Ashraf, I.: Self voting classification

model for online meeting app review sentiment analysis and topic modeling. PeerJ Computer Science **8**, 1141 (2022)

[65] Garg, T., Gupta, S.K.: A hybrid stacked ensemble technique to improve classification accuracy for neurological disorder detection on reddit posts. In: 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 256–260 (2022). https://doi.org/10.1109/CICN56167.2022.10008283

[66] Reshi, A.A., Rustam, F., Aljedaani, W., Shafi, S., Alhossan, A., Alrabiah, Z., Ahmad, A., Alsuwailem, H., Almangour, T.A., Alshammari, M.A., *et al.*: Covid-19 vaccination-related sentiments analysis: A case study using worldwide twitter dataset. In: Healthcare, vol. 10, p. 411 (2022). MDPI

[67] Raman, S., Gupta, V., Nagrath, P., Santosh, K.: Hate and aggression analysis in nlp with explainable ai. International Journal of Pattern Recognition and Artificial Intelligence **36**(15), 2259036 (2022)

[68] Sharma, D., Gupta, V., Singh, V.K., Chakravarthi, B.R.: Stop the hate, spread the hope: An ensemble model for hope speech detection in english and dravidian languages. ACM Transactions on Asian and Low-Resource Language Information Processing (2025)

[69] Bokolo, B.G., Liu, Q.: Advanced comparative analysis of machine learning and transformer models for depression and suicide detection in social media texts. Electronics **13**(20), 3980 (2024)

[70] Kulkarni, S.S., Hareesh, B.V.N., Enduri, M.K., *et al.*: Explainable depression detection in social media using transformer-based models: A comparative analysis of machine learning. In: 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 321–325 (2024). IEEE

[71] Aldhyani, T.H., Alsubari, S.N., Alshebami, A.S., Alkahtani, H., Ahmed, Z.A.: Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. International journal of environmental research and public health **19**(19), 12635 (2022)