Journal of Computational Social Science (2024) 7:2277–2307
https://doi.org/10.1007/s42001-024-00307-1

RESEARCH ARTICLE

Mental disorder and suicidal ideation detection from social media using deep neural networks

Özay Ezerceli[1] · Rahim Dehkharghani[2]

Received: 13 October 2023 / Accepted: 26 June 2024 / Published online: 6 July 2024

Abstract

Depression and suicidal ideation are global reasons for life-threatening injury and death. Mental disorders have increased especially among young people in recent years, and early detection of those cases can prevent suicide attempts. Social media platforms provide users with an anonymous space to interact with others, making them a secure environment to discuss their mental disorders. This paper proposes a solution to detect depression/suicidal ideation using natural language processing and deep learning techniques. We used Transformers and a unique model to train the proposed model and applied it to three different datasets: SuicideDetection, CEASEv2.0, and SWMH. The proposed model is evaluated using the accuracy, precision, recall, and ROC curve. The proposed model outperforms the state-of-the-art in the SuicideDetection and CEASEv2.0 datasets, achieving F1 scores of 0.97 and 0.75, respectively. However, in the SWMH data set, the proposed model is 4% points behind the state-of-the-art precision providing the F1 score of 0.68. In the real world, this project could help psychologists in the early detection of depression and suicidal ideation for a more efficient treatment. The proposed model achieves state-of-the-art performance in two of the three datasets, so they could be used to develop a screening tool that could be used by mental health professionals or individuals to assess their own risk of suicide. This could lead to early intervention and treatment, which could save lives.

Keywords Suicidal ideation detection · Social media content · Word embedding · Deep neural network · BERT transformers

These authors contributed equally to this work.

Extended author information available on the last page of the article

Introduction

ideation is a serious mental disorder condition that may lead to suicide. It refers to persistent thoughts, fantasies, or preoccupations with self-death and self-harm. According to the World Health Organization (WHO), suicide is one of the leading causes of death worldwide, claiming the lives of more than 700,000 people each year [1].

Detecting and addressing suicidal ideation in its early stages is crucial to effective treatment. Traditional approaches to identifying suicidal ideation often rely on self-reporting through clinical interviews or surveys [2]. Despite the high accuracy of these methods, they are limited in their reach, as individuals may be reluctant to reveal their thoughts and feelings in interviews or surveys due to social stigma or fear of being judged.

Detection of suicidal ideation from social media can overcome the limitations of traditional assessment methods. By analyzing publicly shared comments, and interactions of individuals on social media, researchers can gain insight into their mental and emotional states, including signs of distress and indications of suicidal ideation.

There is a growing interest in using machine learning (ML) and deep learning (DL) methods to detect suicidal ideation from social media posts. Deep learning-based algorithms provide promising results in NLP problems. Among deep neural network models, transformer-based algorithms achieve the highest performance; however, the existing research is far from ideal. In this paper, we attempted to improve the state-of-the-art performance achieved in this field by proposing new deep neural networks and transformer-based models and applying them to three benchmark datasets. Specifically, we proposed three different approaches and applied them to three different datasets, namely SuicideDetection [26], CEASEv2.0 [16] and SWMH [23]. Two models were designed for binary classification, but the third one accomplished a 5-class classification where the classes are five different states of being suicidals.

The contributions of the current work can be summarized as follows.

- This paper proposes three different deep neural network architectures each of which was designed and specialized for one of the following datasets: SuicideDetection, CEASEv2.0, and SWMH.
- The proposed models achieve state-of-the-art performance in two datasets (Suicide Detection, CEASEv2.0), outperforming previous models by 0.02 and 0.01 points in terms of the F1 score.
- This paper compares classic machine learning with deep learning techniques when applied to the depression and suicidal ideation detection problem.

In the remainder of this paper, Sect. "Literature review" reports outstanding previous work on suicidal/depression detection. A detailed explanation of the proposed approach is provided in Sect. "Proposed approach", which is followed by experimental evaluation in Sect. "Experimental evaluations". Results are discussed in

Sect. "Discussion", and conclusions and future work are provided in Sect. "Conclusion and future work".

Literature review

Suicidal ideation can occur in different manners and various aspects. Identification of this intention from social media has some limitations such as the informal language of the posts, and the demographic characteristics of the users which makes this text classification task challenging. These challenges are tackled by a group of researchers in the literature using an appropriate feature list. In their recent study, Shah et al. [43] proposed a hybrid feature extraction approach that combines Genetic algorithm and Linear Forward Selection (LFS) methods to select the most relevant linguistic and computational features for suicidal ideation detection in social media, and applied the proposed approach to 7098 numbers of Reddit's SuicideWatch subReddit social media posts as their dataset. The article compared various classification methods using different feature sets and concluded that Random Forest achieves the highest performance among others.

Researchers often do suicidal ideation classifications as binary classification or multiclass classification. Unlike most of them, [7] discusses the use of multitask learning (MTL) in a deep learning framework [52] to estimate the risk of suicide and mental health using a union of multiple Twitter datasets that are manually annotated. The authors compare their MTL model to a well-tuned single-task baseline to predict a potential suicide attempt and the presence of atypical mental health with AUC> 0.8. The final dataset contains 9,611 users in total.

Another study [16] is one of the first examples of multitask classification for the extension of suicide notes (CEASE dataset [15]) which is a total of 315 real-life suicide notes. The proposed multitask framework uses GloVe embedding and the Bi-GRU DL layer to detect depression, sentiment, and multi-label emotion. They achieved 0.74 accuracy on the depression detection task.

Deep neural networks have shown high performance in various tasks such as image classification, natural language processing, and speech recognition. In many text classification tasks, deep learning models outperform classic machine learning models and traditional data analysis approaches [20].

In [46], the authors propose a deep neural network model that combines Long Short-Term Memory (LSTM) and convolutional neural networks (CNN) that are pre-trained with 300-dimensional word2vec word embedding vectors to detect suicidal ideation on Reddit social media dataset [21]. The dataset is created by 3549 suicidal and 3652 non-suicidal posts. Various NLP techniques, such as TF-IDF, BOW, and statistical features, are employed to encode words and extract features from text data. The authors observed evidence of hopelessness, frustration, anxiety, and signs of loneliness that point to suicidal ideation. The proposed combined neural network model outperformed other approaches in the literature when applied to the Reddit dataset [21] with a 0.93 F1 score.

Another study [5] proposed a DL model based on CNN-BiLSTM.The CNN-BiLSTM model achieved 0.95 suicidal ideation detection accuracy by using

textual features. The authors suggested that Reddit users at risk of suicide may have mental and psychological health problems. Our proposed approach could achieve 0.97 as an F1 score on the same dataset and outperformed pioneering approaches in the literature.

Some users might tend to hide their real identities and feelings and not explicitly express their feelings on social media due to privacy issues. This situation leads to the trade-off between privacy and prevention [12] because it prevents reaching a fraction of users that might be at risk; however, those users might express their feelings and intentions implicitly using abstract or sarcastic sentences such as "What a life!", "Ohh I am so lucky!". Such sentences in social media could be misleading and difficult to understand the intention of their writer, but generally, suicide notes are written shortly, which can be used as clues to detect depressed and in-danger users before suicide. Taking into account these limitations, the authors in [16] provide a new corpus of suicide notes in English, named CEASE, which has been created and annotated with 15 fine-grained emotion labels (forgiveness, peace, and happiness, love, pride, hopefulness, thankfulness, blame, anger, fear, abuse, sorrow, hopelessness, guilt, information, instructions). This corpus consists of 2393 sentences from about 205 suicide notes collected from various sources. Deep learning-based ensemble models of CNN, the gated recurrent unit (GRU), and LSTM models are used to detect emotions in the corpus with an accuracy of about 0.60.

Identifying the meaning of the context is a crucial part of understanding the ideation [32]. Most of the time, context composes various parameters that might affect the context such as sentence length, words starting with capital letters, special keywords, special stopwords, negative meanings, topic [17], abbreviations, and punctuation marks.

In [32], the authors propose a probabilistic framework that models users' online activities as a sequence of psychological states over time and estimates emotional states by incorporating context history using Conditional Random Fields (CRF).

In addition, some research works [19, 39] have investigated statistical differences between textual features (Linguistic Inquiry and Word Count (LIWC) [35], Bag of Words (BoW) [38] or n-grams [27] and word embeddings [28]) and behavioral characteristics of each risk group for suicidal ideation. They evaluated statistical and deep-learning-based approaches to handling multimodal data for suicidal ideation detection of users from Twitter. The obtained results contribute to our understanding of how the combination of textual, visual, relational, and behavioral data outperforms each model in isolation by 0.08.

Despite interest in context, advances in deep learning and attention mechanisms of neural networks in the NLP tasks lead to better representing of the user's context and to capturing the relationship between sentences[51]. In [22], the authors propose a model for effectively encoding relational text to detect suicidal ideation and mental disorders. The attention mechanism is incorporated to prioritize crucial relational features. The datasets used in this study include the UMD Reddit Suicidality Dataset [44] from 11,129 users who posted on SuicideWatch, and 11,129 users who did not, and also the SuicideWatch and Mental Health Collection (SWMH) dataset with a total of 54,412 posts. The authors report F1 scores of 0.54 and 0.64, respectively for the UMD and SWMH datasets.

The proposed ideas in [22] are supported by [6] which proposes a hybrid and ensemble method consisting of LSTM and LR. This method was tested on three different datasets: CLPsych 2015 [11], Reddit [37], and eRisk[30]. The authors concluded that word embeddings can be one of the driving sources of performance differences among hybrid models and DL with the model. Their best performance was achieved on the Reddit dataset with 0.77 as the F1 score.

Transformers as recent advances in the NLP field have made a wide variety of tasks more accurate and faster, as they provide a better understanding of sequence data and a shorter training time because they use larger dimensions for word and sentence embeddings [47] and process the whole input in parallel.

Since transformers are an open-source library, researchers can extend them to build more advanced architectures [49]. Pre-trained transformers with domain-specific data can be used for specific tasks. In [24], the authors developed two pre-trained masked language models, MentalBERT and MentalRoBERTa, for the mental healthcare research community. The authors used the language representations pre-trained in the target domain, for improving the performance of mental health detection tasks. They evaluated these models and several variants of pre-trained language models and could achieve a recall of 0.70 and an F1 score of 0.72 by the MentalRoBERTa model.

In another work [42], the authors propose a time-aware transformer-based model called STATENet for preliminary screening of suicidal risk on social media. The model jointly learns from the language of the tweet and the emotional historical spectrum in a time-sensitive manner to detect users with suicide ideation. The authors use Twitter timeline data from the dataset proposed by [45] which contains 34,306 tweets. They report an F1 score of 0.81 for STATENet, which is greater than the F1 score of 0.77 reported by the best-performing baseline model. The authors concluded that suicide risk exists in a diverse spectrum and that simplification of binary labels could lead to artificial notions of risk.

Proposed approach

Data availability is an important factor in choosing a classification model for text classification problems. Three different datasets with different characteristics have been used in the current research. We have presented three unique models in this work, which are applied to all three datasets taking into account their diverse features. The framework of the proposed models is illustrated in Fig. 1. Firstly, we applied different pre-processing steps to all datasets. Then, the sequence data were converted to vectors with the help of different word embedding models such as Word2Vec, fastText, and GloVe. Different classic machine and deep learning algorithms have been used in the architecture of the proposed models. Specifically, the BiLSTM layer, the attention layer [47], and different types of transformers (BERT, RoBERTa) are used along with the additional dropout and batch normalization layers within the proposed models. Two out of three models result in binary labels (suicidal, non-suicidal), while the third one results in multi-class labels (depression, suicidewatch, anxiety, offmychest, bipolar). Therefore, the sigmoid activation functions

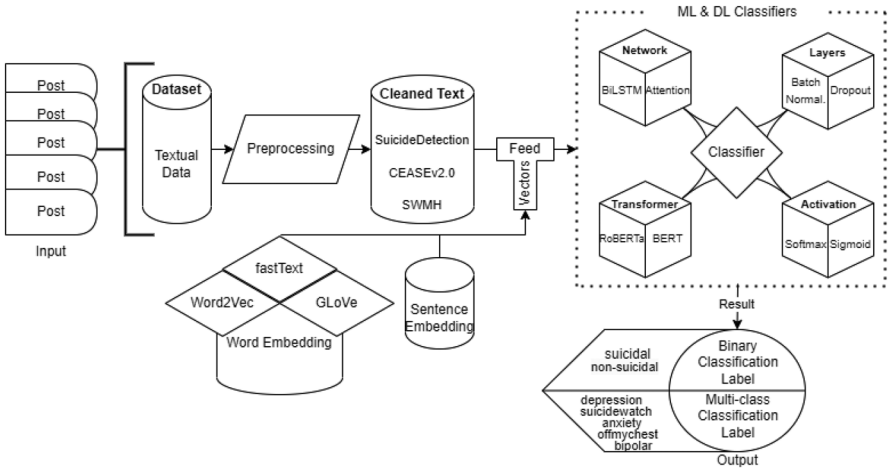

Fig. 1 Framework of the proposed suicidal ideation detection classifier

[34] and the binary cross-entropy loss function are used for the SuicideDetection and CEASEv2.0 datasets, but the softmax activation function [9] and the sparse categorical cross-entropy loss function are used for the SWMH dataset. All steps in Fig. 1 are explained with details in the following sections (see Table 1).

Preprocessing

Textual datasets are mostly unstructured because they are written in natural language by totally different people. Social media posts often contain images, videos, emoticons, and other multimedia elements. These data should be converted to structured form by applying a preprocessing step before they are fed into the model as input. In the preprocessing step, we apply a set of methods on the unstandardized dataset using common NLP libraries such as NLTK [29], Beautiful soup [41] and Neattext [4] to

- Reduce noise in the data to achieve higher performance;
- Reduce the size of the data for more efficiency;
- Remove irrelevant features for greater consistency.

Some steps in the preprocessing phase are commonly applied to all datasets, but others are dataset-specific. Table 4 shows which pre-processing step has been applied to which dataset.

We have collected a corpus called a contraction dictionary from the social media posts, which is composed of 151 most common contractions and their openings (i.e., "don't" -> "do not", "ain't" -> "is not"). We removed HTML codes, links, new lines,

Table 1 Details of DL models for SuicideDetection dataset. (Each column represents a different set of parameters for a model of the dataset

Layer & Parameters	Models for SuicideDetection Dataset								
Data Split	%80 Training - %20 Test								
Input Shape	75	100	100	191	191	188	200	184	184
Embedding	GloVe				FastText		BERT	Sentence	Sentence
Embedding Model	glove.840B.300d				crawl-300d-2 M-subword		Bert-base-uncased	All-MiniLM-L6-v2	All-mpnet-base-v2
Embedding Size	300	300	300	300	300	300	768	384	768
Network	BiLSTM		2BiLSTM						
Activation	Relu	Relu	Relu	Relu	Relu	Relu		Relu	Relu
Dropout Rate	0.5	0.5	0.5	0.7	0.5	0.3		0.3	0.3
Flatten	✓	✓	✓	✓	✓	–	✓	✓	✓
Batch normalization	–	–	–	✓	✓	✓		✓	✓
Epoch	10	5	5	15	10	20	2	20	20
Batch	256	256	128	256	128	128	64	128	128
Optimizer	RMSprop	Adam							

extra whitespace, and special characters. We removed stopwords except for the keyword "I" because it emphasizes suicidal ideation.

Preprocessing steps are separately applied to datasets according to their nature and writing style; for example, due to spelling errors in the CEASE dataset, spell correction is applied only on this dataset, or due to the short length of sentences in the SWMH dataset, singular to plural form conversion is not applied on this dataset.

Table 3 lists examples of three different sentences from each dataset before and after preprocessing.

Word embedding

Word embedding is a technique used in natural language processing (NLP) to represent words as numeric vectors in a high-dimensional space (100, 200, etc.). This allows words to be compared with each other based on their semantic meaning; e.g., words with similar meanings will have similar vectors. We used the most popular word embeddings, namely, GloVe, and fastText, and also transformer-based sentence embeddings for training the proposed models.

GloVe: Global Vectors For Word Representation

GloVe [36], as one of the prominent word embedding models, unlike Word2vec [31], utilizes global statistics (word co-occurrence) along with local statistics (local context information of words) to generate word representations. This enables the identification of meaningful semantic relationships among words. We started our experiments by using the GloVe embeddings first. Equation (1) shows the cost function for GloVe word embeddings.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - logX_{ij})^2 \tag{1}$$

where:

$\mathbf{X_{ij}}$ represents how often word i appears in the context of word j

$\mathbf{w_i}$ the vector for of the main word

$\mathbf{\tilde{w}_j}$ represents the vector for the context word

$\mathbf{b_i}$, $\mathbf{\tilde{b}_j}$ is the scalar biases for the main and context words

and **f** is the weighting function that helps us to prevent learning only from extremely common word pairs

FastText

FastText [8] is an open-source library for efficient learning of word representations. It provides two models for the computation of word representations: skip-gram and continuous-bag-of-words (cbow). The skip-gram model acquires knowledge to

Table 2 Details of models for CEASEv2.0 dataset. (Each column represents a different set of parameters for a model of the dataset.)

Layer & Parameters	Models for CEASEv2.0 Dataset	
Data Split	%80 Training - %20 Test	
Input Shape	44	43
Embedding	Sentence	fastText
Embedding Model	All-mpnet-base-v2	crawl-300d-2 M-subword
Embedding Size	768	300
Network	2BiLSTM	BiLSTM
Activation	Softplus	relu
Dropout Rate	0.18	0.25
Flatten	–	–
Batch Normalization	✓	✓
Epoch	42	10
Batch	265	32
Optimizer	Nadam	Adam
Callbacks		
Early Stopping	✓	✓
Reduce LR	✓	✓
Checkpointer	–	–
F1-score (%)	68	**75**
AUC Score (%)	–	**70**

estimate the neighbors of a target word, while the cbow model forecasts the target word based on its neighbors. Using subword-level information, it builds vectors for unknown words. It uses cosine similarity between the vectors. Equation (2) shows that cosine similarity can be calculated as the dot product of the vectors normalized by their size.

$$cosine\_similarity(u, v) = \frac{u \cdot v}{||u|| \cdot ||v||} \tag{2}$$

Transformer-based sentence embedding

SentenceTransformers [40] are state-of-the-art sentence embeddings that use siamese and triplet network structures. In [40], authors concluded that Sentence-BERT (SBERT) structure which uses siamese and triplet network structures, is 46.8k faster than cosine similarity to find the most similar pair. Transformers have their own preprocessing step because, unlike other ML & DL algorithms, transformer models such as BERT, RoBERTa have their own input type. While the granularity in Word2Vec and fastText is at the word level, sentence embeddings work at the sentence granularity level. There is a collection of pre-trained models that have been

Table 3 Example preprocessing for each dataset

Dataset	Before	After	Label
SuicideDetection	It ends tonight.I can't do it anymore. I quit	It ends tonight I anymore I quit	Suicide
	I'm f*cked assignment is due tomorrow and I haven't even started yet	I fucked assignment tomorrow I started	Non-suicide
	PLEASE HELP ME I CANT STOP SCREAMING I NEED HELP	PLEASE HELP ME I CANT STOP SCREAMING I NEED HELP	Suicide
CEASEv2.0	But you shattered my dreams	Shattered dream	Depression
	I expect pain very likely to outweigh happiness and satisfaction in my life	I expect pain likely outweigh happiness satisfaction life	Depression
	I love you completely you will find my body on the lot on the north side of the house	I love completely body lot north house	Non-depression
SWMH	I just took about 30 anxiety pills. I doubt it will kill me but if I does, great!	Take anxiety pills doubt kill great	Depression
	I'm gonna do it tonight I've attempted 4 times already. Why not once more?	Go tonight attempt times already	Suicidewatch
	I fell in love while travelling, and ruined it. I'm never going to see him again and it's wrecking me	Fall love travel ruin never go see wreck	Offmychest

Table 4 Preprocessing steps for each of the datasets

Preprocessing	Datasets		
	Suicid-eDetection	CEASEV2.0	SWMH
Remove HTML	✓	✓	✓
Remove new lines	✓	✓	✓
Remove Links	✓	✓	✓
Remove special characters	✓	✓	✓
Replacing contractions	✓	✓	✓
Remove stopwords	✓	✓	✓
Words to singular	✓	✓	–
Spell correction	–	✓	–
Lemmatization	–	–	✓

fine-tuned for various tasks. The embedding models that we used for the SuicideDetection and CEASEv2.0 datasets can be found in Tables 1 and 2.

Classification

We have built three different classification models; Two of them accomplish a binary classification and the third one does a multi-class classification. In the training phase, we used "binary-crossentropy" which is calculated in equation (3) for the binary classification, and "SparseCategoricalCrossentropy" for the multi-class classification. The estimated labels for the SuicideDetection and CEASEv2.0 datasets are either “suicidal” or “non-suicidal” but for the SWMH dataset it is one of the five classes: “depression”, “suicidewatch”, “anxiety”, “offmychest”, and “bipolar”. These class labels are also shown in Fig. 1.

$$BCE(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{3}$$

$$SCCE(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{4}$$

where:

BCE is Binary Crossentropy

SCCE is Sparse Categorical Crossentropy

N is the output size

y is the true label value

and $\hat{y}$ is the predicted label value

In the architecture of the proposed models, various state-of-the-art word and sentence embedding layers, and neural network layers have been used.

Hyperparameter optimization is applied to the parameters of these layers and also to the general parameters of the model. Different optimizers such as Root Mean Square Propagation (RMSprop), Adam [25], and Nadam [14] have been used. The BERT model also uses an unmodified Adam optimizer. In all datasets, the Adam optimizer was more successful in finding global minima for the current problem since it uses a combination of two gradient descent methodologies: momentum and RMSprop.

We have used Rectified Linear Units (ReLU) [3] activation function as shown in equation (5) primarily along with the softplus function which is a smoothed version of ReLU as shown in equation (6) as the activation function of the dense layers.

$$f(x) = max(0, x) \tag{5}$$

$$f(x) = log(1 + e^{x}) \tag{6}$$

Tanh activation function [33] as shown in equation (7) for BiLSTM layers with the relu function for the dense layers usually obtains promising results.

$$f(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x}) \tag{7}$$

Activation of the last dense layers differs in the proposed models as two models use the sigmoid function as in equation (8) for the binary classification while the third one uses the softmax function as in equation (9) for multi-class classification.

$$f(x) = 1/(1 + e^{-x}) \tag{8}$$

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{9}$$

Overfitting is a trap that trained models may fall into it. In this trap, while a model performs very well when applied to the train set, it shows a very low performance on the test set (unseen data). In order to prevent this issue and generalize the proposed model, we used the dropout method with different dropout rates and batch normalization layers. The dropout layers randomly drop neurons from the network during training, while the batch normalization layers normalize the input data to each layer in order to reduce the internal covariate shift. For SuicideDetection dataset we used the rate between 0.3−0.7, where 0.3 dropout rate achieved the best results. For the CEASEv2.0 dataset, this rate was between 0.18 and 0.25 and the 0.25 dropout rate achieved the best results. This difference is due to the dataset size; SuicideDetection dataset is 46.4k times larger than the CEASEv2.0 dataset.

Callbacks are utilities in machine learning to prevent overfitting. We used Early Stopping, Reducing the LR, and Checkpointer methods. Early stopping stops training after a specific number of attempts if there is no further advancement in

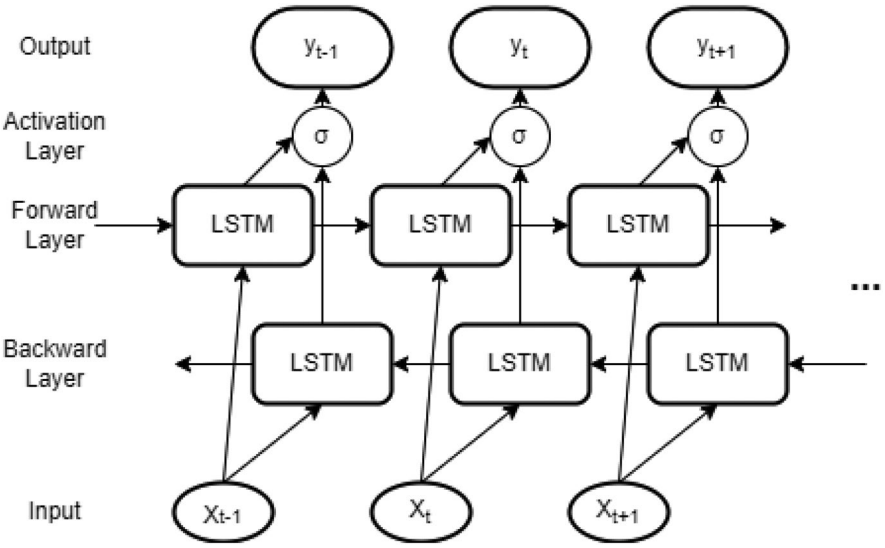

Fig. 2 BiLSTM network stucture

validation accuracy. Therefore, there would be no waste of time for training. Reducing the LR improves the accuracy of a model by allowing it to adjust the learning rate during training. Checkpointer callback saves the model weights at the end of each epoch by watching the performance of the model so that we can save the best model if there is any interruption or crush.

BiLSTM networks

Our proposed models for the SuicideDetection and CESEv2.0 datasets use BiLSTM [18] network. BiLSTM layers have the advantage of capturing both the forward and backward directions of the given context as shown in Fig. 2, allowing for a better understanding of the sequence as a whole. Figure 3 shows the proposed model architecture for SuicideDetection dataset. We used two layers of BiLSTM; For the first layer, a parameter called return_sequences is set to true because return_sequences returns the hidden state output for each input time step, and the output of each time step is sent to the second BiLSTM layer.

CEASEv2.0 dataset has the shortest user comments among other datasets and the best model for this dataset has 43 input shapes and is trained for 10 epochs with 16 batch size. The model also has a batch normalization layer before the fully connected output layer. The model architecture is shown in Fig. 4.

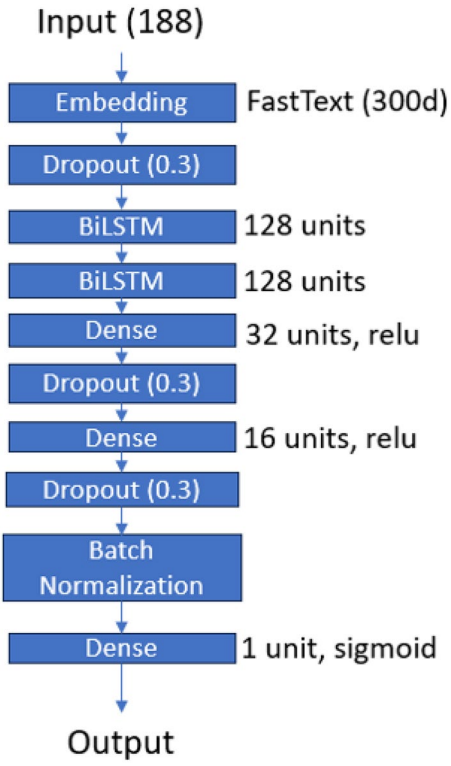

Fig. 3 Proposed model architecture for suicidedetection dataset

BERT transformer

We used the BERT Transformer [13] for the SWMH dataset. There exists a multi-head attention mechanism in each block of the transformers. The attention mechanism allows the extraction of only the most relevant information from a given input, which reduces the computational complexity of a model [10]. The proposed model's architecture is shown in Fig. 6. The model trained for the SWMH dataset has 256 input shapes and is trained for 2 epochs. The total value of the trainable parameters is 108.706.565. The BERT itself has 768 dimensions. The summary of the SWMH model is given in Fig. 5.

State-of-the-art deep neural network-based algorithms such as transformers provide higher efficiency than traditional machine learning algorithms because they process large amounts of data in parallel and can automatically extract features from data without relying on hand-crafted features.

We have trained our models with different input shape sizes. This value depends on the maximum text length of each document because word plays its role as a feature, and using all words as features results in better performance. Different pre-processing steps may result in different input shapes. The best model for the SuicideDetection dataset has 188 input shapes and is trained for

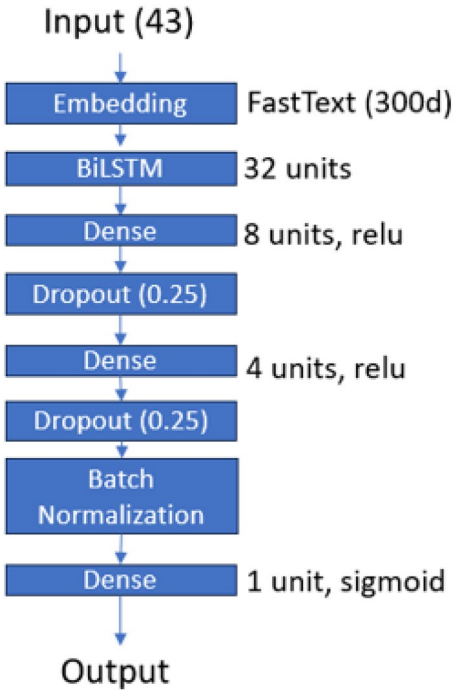

Fig. 4 Proposed model architecture for CEASEv2.0 dataset

Fig. 5 SWMH model summary

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_ids (InputLayer)      [(None, 256)]             0

 attention_mask (InputLayer) [(None, 256)]             0

 bert (TFBertMainLayer)      TFBaseModelOutputWi       108310272
                             thPoolingAndCrossAt
                             tentions(last_hidde
                             n_state=(None, 256,
                              768),
                              pooler_output=(Non
                             e, 768),
                              past_key_values=No
                             ne, hidden_states=N
                             one, attentions=Non
                             e, cross_attentions
                             =None)

 intermediate_layer (Dense)  (None, 512)               393728

 output_layer (Dense)        (None, 5)                 2565

=================================================================
Total params: 108,706,565
Trainable params: 108,706,565
Non-trainable params: 0
```

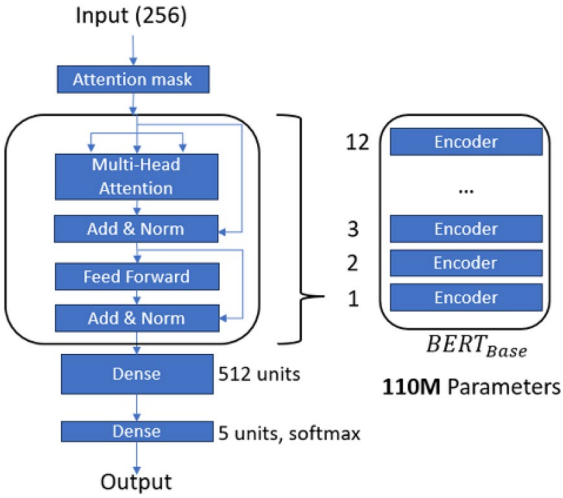

Fig. 6 Proposed model architecture for SWMH dataset

20 epochs with 128 batch sizes. Note that here the best model means the model with the best performance, e.g., the highest accuracy.

Experimental evaluations

Dataset

In this section, a comprehensive evaluation of the proposed method is provided. The details of the datasets, the evaluation metrics, and the results, followed by a discussion of those results are given in the following subsections.

Suicide detection dataset

The Suicide Detection Dataset[1] [26] is a collection of user posts on Reddit ("SuicideWatch" subreddit) that are publicly available on Kaggle. The dataset consists of 232,074 posts on 'SuicideWatch' from December 16, 2008 to January 2, 2021. It has 116,037 suicide posts and 116,037 non-suicidal posts. The SuicideWatch subreddit refers to a monitoring procedure designed to prevent suicide attempts by individuals who display suicidal warning signals since they can be at risk for intentional self-harm.

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

CEASEv2.0 dataset

The CEASEv2.0 dataset [16] is the extended version of CEASE dataset [15] which is annotated with 15 fine-grained emotions at the sentence level of suicide notes in English, comprising 2393 sentences (from 205 suicide notes). With the addition of 2539 sentences to the base version, the dataset consisted of 4932 sentences collected from a totally collected 325 suicide notes. CEASEv2.0 is publicly available for research purposes. It is the most complex/challenging dataset among these three datasets because of the smaller data size and an unbalanced rate of labels.

SWMH dataset

The SuicideWatch and Mental Health Collection called the SWMH dataset [23] is the collection of a total of 54,412 posts specific to the subreddits of depression, suicidewatch, anxiety, offmychest, and bipolar using the Reddit API. Unlike the other datasets we used for binary text classification, we used this dataset for multi-class classification.

Evaluation metrics

In our experiments, we evaluated the efficiency of the proposed methods based on the AUC score and the F1 score. The AUC score is the area under the ROC curve, which measures the model's ability to distinguish between positive and negative classes as shown in equation (10). The F1 score is the harmonic mean of precision and recall as in (13), which is a measure of the model's ability to correctly classify positive and negative classes. The calculation of precision and recall is mentioned in (11) and (12) respectively. Both the AUC and the F1 scores can be used instead of recall and precision metrics. Since we evaluated multi-class classification for the SWMH dataset, the AUC score for multi-class classification is calculated by taking the average of the AUC scores for each class.

$$AUC = \frac{1}{n}\sum_{i=1}^{n}(TP_i + \frac{1}{2}FN_i) \tag{10}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{11}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{12}$$

$$F1 = 2 * (precision * recall)/(precision + recall) \tag{13}$$

Table 5 Best proposed models for each of the datasets

Dataset	Size	Description	Method	Result
Suicide detection [26]	232.074	Reddit data (%50-%50)	FastText, BiLSTM	0.97 F1-score, 0.996 AUC score
CEASEv2.0 [16]	4.932	%64 non-depresision, %36 depression	FastText, BiLSTM, word correction, initial bias	0.75 F1-score, 0.70 AUC score
SWMH[a] [23]	54.410	Reddit data (multi-class)	BERT	0.68 F1-score

[a] SWMH has 5 classes

Table 6 Comparison and evaluation of ML models for suicidedetection dataset

ML Algorithms	Score (%)			
	Accuracy	Precision	Recall	F1 Score
MNB[b]	56.32	16.42	78.52	27.16
RF[c]	81.71	82.23	81.14	81.68
LR[d]	70.78	71.53	70.78	70.88
SVM[e]	93.21	93.31	93.21	**93.21**
DT[f]	82.97	83.94	82.97	83

[b]Multinomial Naive Bayes (MNB) [c]Random Forest (RF) [d]Linear Regression (LR) [e]Support Vector Machine (SVM) [f]Decision Tree (DT)

Table 7 Comparison and evaluation of ML models for SWMH dataset

ML Algorithms	F1-Score (%)				
	One-hot Encoding	TF-IDF	TF-IDF ngrams	TF-IDF char ngram	FastText
MNB[g]	62	58	52	57	–
LR[h]	62	**67**	53	65	–
k-NN[i]	45	52	25	49	–
RF[j]	58	62	50	59	–
SD[k]	60	66	52	65	21
GB[l]	60	61	46	62	35
XGBoost	66	66	53	65	40

[g]Multinomial Naive Bayes (MNB) [h]Linear Regression (LR) [i]K-Nearest Neighbors (k-NN) [j]Random Forest (RF) [k]Stochastic Descent(SD) [l]Gradient Boosting (GB)

Results and comparison

The SuicideDetection dataset is experimented with the ML and DL algorithms as

Table 8 Comparison and evaluation of DL models for SWMH dataset

DL Algorithms	FastText word embedding	
	F1-score (%)	AUC Score (%)
Shallow network	66	86.8
DL (just embedding)	64	86.3
DL (Emb.+2 Dense)	63	86.4
RNN	55	–
CNN	60	–
LSTM	65	86.5
CNN + LSTM	64	87.3
CNN + GRU	63	85.8
RNN + GRU	64	85.3
2BiLSTM	62	84.2
Bidirectional GRU	64	85.0
RCNN	66	87.5
RCNN-v2	66	**87.8**
RCNN-v3	65	86.5
BERT	**68**	–

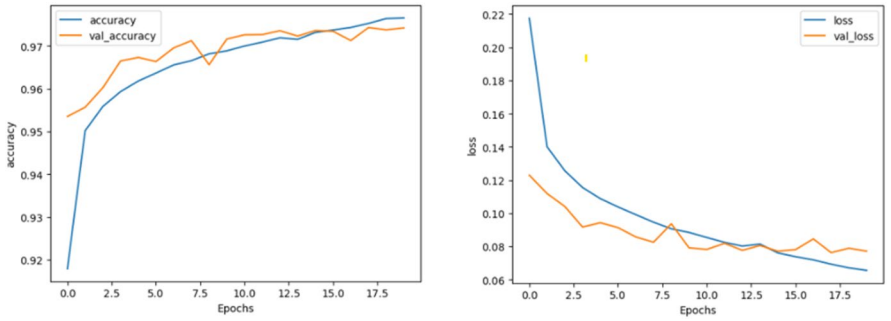

Fig. 7 Accuracy & loss graph of model for suicidedetection

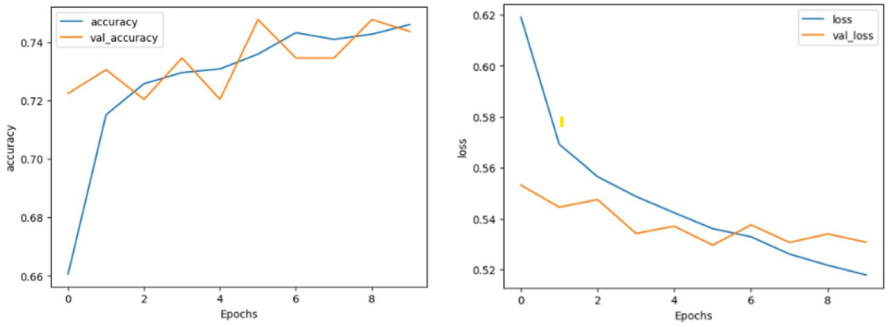

Fig. 8 Accuracy & loss graph of model for CEASEv2.0

shown in Tables 6 and 1. The best model for the Suicide Detection dataset uses deep neural network models. The designed model consists of ten layers with different parameter values as in Fig. 3 where the order of layers is as follows: Embedding + Dropout (0.3) + 2BiLSTM (128 units) + Dense (32 units) + Dropout (0.3) + Dense (16 units) + Dropout (0.3) + Batch Normalization + Dense (1 unit). FastText (crawl-300d-2 M-subword.bin) is used to generate word embedding vectors. The model achieved a 0.97 F1 score and 0.9742 accuracy as a result of training for 20 epochs with 128 batch sizes on the SuicideDetection dataset. The proposed model outperformed the state-of-the-art model [5] achieving 0.95 accuracy by using CNN-BiLSTM with Word2Vec. The AUC score of our model on the SuicideDetection dataset is 0.996. Besides, our validation loss is 0.08 as in Fig. 4, while in [5], the validation loss of 0.15.

The CEASE dataset is the most difficult dataset to classify among the three due to its small size and unbalanced classes. This small size and the short length of its posts highlight the role of spell correction in this dataset, as we do not want to lose any words due to syntax errors in a short post. The designed model for this dataset includes eight layers with different parameters shown in Fig. 4. The order of layers is as follows: Embedding + BiLSTM (32 units) + Dense (8 units) + Dropout (0.25) + Dense (4 units) + Dropout (0.25) + Batch Normalization + Dense (1 unit). Due to its small size, the designed model for the dataset is smaller than the others in terms of layer number and neuron number in each layer. The model is trained for 10 epochs with 16 batch sizes, with a bias initializer which is used due to unbalanced classes (Fig. 7). This model achieved a 0.75 F1 score as illustrated in Fig. 8 and a 0.70 AUC score as shown in Table 5. The achieved F1 score outperformed the best model of [16] with 0.7435 as the F1 score using the GLoVe + Bi-GRU approach.

The SWMH dataset had been tested with ML algorithms in Table 7 and DL algorithms in Table 8. The best F1 score is achieved by using a BERT transformer trained for two epochs with batch size 16 which resulted in 0.68 as the F1 score. The model outperformed the proposed model in [23] which is built with RN and achieves 0.64 F1 score. The state-of-the-art F1 score for this dataset is 0.72 using the MentalRoBERTa model as in [24]. The intuition behind this success is that this transformer has been fine-tuned with domain-specific data–mental health-related posts collected from Reddit.

We have designed and tested different models, the details of which are listed in Tables 1, 2, and 6, 7, 8. At the end of the day, we selected the best model for each dataset which is shown with details in Table 5.

Discussion

A discussion of data and obtained results, as well as error analysis, the limitations of the proposed models, and practical and ethical considerations for suicidal ideation detection models are given in this section.

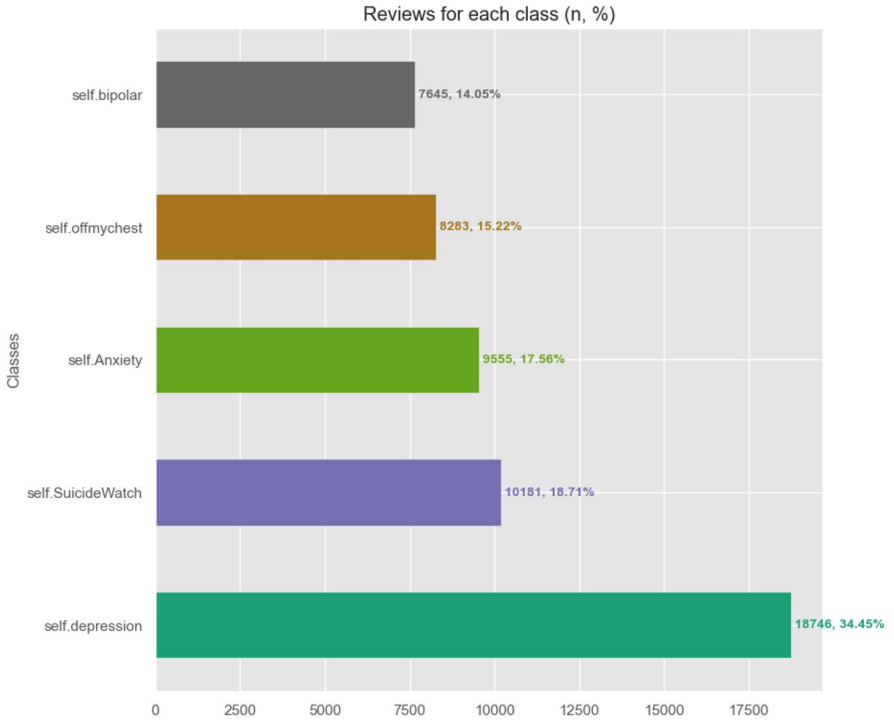

Fig. 9 Class rates of SWMH dataset

Data analysis

Data is probably the first factor that affects the model performance as NLP applications are generally data-driven. SuicideDetection dataset has a balanced rate of suicidal and non-suicidal data. The SWMH dataset has five different classes, and the weight of each class is calculated using the compute_class_weight method in the Sklearn library as shown in (14). Class weights labels are presented in Fig. 13. The ratio of the dataset is balanced (ratio > 0.1) with the value of 0.408 (depression label=0.5805) / 1.4234 (weight of the bipolar label)) as calculated in (15). Also, the distribution of class rates of SWMH can be seen in Fig. 9.

$$w_j = \frac{n\_samples}{n\_classes * n\_samples_j} \tag{14}$$

where:

w_j is the weight for each class (j signifies the class)
$n\_samples$ is the total number of samples or rows in the dataset
$n\_classes$ is the total number of unique classes in the target
$n\_samples_j$ is the total number of rows of the respective class

Fig. 10 WordCloud of all datasets combined

word_count	11.055524	word_count	7.978964
sent_count	1.012593	sent_count	0.987441
avg_sentlength	10.979870	avg_sentlength	8.070517
unique_word_count	10.137951	unique_word_count	7.442700

Fig. 11 Sentence features of CEASEv2.0 dataset (depression - non-depression)

word_count	138.657230	word_count	43.125460
sent_count	6.984608	sent_count	2.384041
avg_sentlength	46.962059	avg_sentlength	23.239315
unique_word_count	87.921921	unique_word_count	30.619035

Fig. 12 Sentence features of suicidedetection dataset (suicidal - non-suicidal)

Fig. 13 Class Weights of SWMH dataset

```
Class weight: 1.1389    class: self.Anxiety
Class weight: 1.0688    class: self.SuicideWatch
Class weight: 1.4234    class: self.bipolar
Class weight: 0.5805    class: self.depression
Class weight: 1.3139    class: self.offmychest
```

$$ratio = \frac{np.min(df.label.value\_counts())}{np.max(df.label.value\_counts())} \tag{15}$$

The CEASEv2.0 dataset is an imbalanced dataset, with 64% non-depression data and 36% depression data. Due to this imbalance, it is necessary to add extra steps, such as initial bias settings, when training the dataset.

The training process may vary depending on the size of the data. In the case of a large dataset, it is necessary to use a deeper network to extract an optimal set of features. This is because a larger dataset (in terms of the number and length of posts) may contain more complex features that a shallower network may not be able to recognize.

A comparison of the performance of the ML algorithms is provided in Table 7 and Table 6 for the SWMH and SuicideDetection datasets. Word embedding techniques such as Fasttext combined with DL algorithms achieve better performance than other word embeddings. This improvement in the performance can be seen in

Table 9 Similarity values of different embedding techniques

Word Pairs	Word2Vec	FastText	Sentence Embedding
(Suicide-Happy)	0.20	**0.14**	0.32
(Family-Love)	0.44	0.39	0.41
(Suicide-Forgive)	0.26	**0.21**	0.26
(Understand-Forgive)	0.52	0.50	0.34
(Suicide-Death)	0.67	0.65	0.75
(Suicide-Life)	0.39	0.41	0.57
(Love-Life)	0.59	0.44	0.48
(End-Suicide)	0.32	0.30	0.27
(Time-suicide)	0.22	0.18	0.31

Table 1, Table 2, and Table 8. This combination explains why our models outperform others [5, 16] for the SuicideDetection and CEASEv2.0 datasets.

The model for the SuicideDetection dataset has ten layers, while the BERT Transformer model for SWMH has twelve encoder blocks. The shallowest model for CEASEv2.0 has seven layers.

Some words appear more than others in users' comments and carry more importance. These keywords such as 'feel', 'life', 'want', 'know", 'think', 'time', and 'help' are shown in Fig. 10 as a word cloud which is created from the combination of the datasets give a general idea of an individual's intentions when thinking about their mental disorders. Among the most 100 frequent words in the datasets, we found words related to self-harm, hopelessness, sadness, and fear, indicating a negative and stressful mental state among some of the users.

Individuals who struggle with suicidal thoughts tend to seek ways to express themselves and connect with others, while those without such thoughts may not have the same motivation to seek support or share their struggling thoughts. This assumption is also supported in Figs. 11 and 12. These figures show that suicidal texts have greater word count, sentence count, average sentence length, and unique word count when compared to non-suicidal texts.

We have used three different embedding techniques. Word2Vec and FastText are based on word embeddings, and sentence embedding is based on transformers. In Table 9, similarity values for some similar and unsimilar word pairs are shown according to three embedding models. The models used for these embedding techniques are glove.840B.300d.pkl, crawl-300d-2 M-subword.bin, and all-mpnet-base-v2 respectively for Word2Vec, FastText and Sentence Embedding. For instance, the word pair suicide-happy has similarity values of 0.20, 0.14, and 0.32 respectively from Word2vec, FastText, and Sentence Embedding. We expect a small value for this word pair since these two words are not semantically similar or related. FastText gives the minimum similarity score also for the word pair suicide-forgive which is 0.21 while the other embedding's similarity scores is 0.26. It can be seen from the values in Table 9 and the performance of proposed models that FastText is more stable and its embedding score are more meaningful.

Table 10 A few misclassified examples from the suicideDetection dataset

Text	Predicted label	True label
Tell story wanted filler filler filler fillller	Suicide	Non-suicide
Start rich start company significant 16 afford big van trip friend time left start working	Suicide	Non-suicide
Poem haiku umaleegamedev hi hello hello stop fucking saying hello know live	Suicide	Non-suicide
Like think unfair life sit like seriously sitting world suck life suck ugh	Suicide	Non-suicide
Fking funny comforting comforting guy love comforting join	Non-suicide	Suicide
Happy suicidei thinking suicide honestly idea pretty good life happy majority time happy person happy life thought especially 2 week thinking ending feeling confused day busy actively stuff fine moment free time click feel lonely sad uninspired overall depressed suggestion lifeadvice know talking people obvious solution trusting people stubborn individual	Non-suicide	Suicide
Optimistici consider optimist best world beautify love life alive universe star sun water great thing amazing look people glad exist know home money going college nice town load friend group activity work community constantly fulfilling project eat good food drink pleanty water exercise daily life better thats want kill feel happiness concept unimaginable family disgusted father spoken month cut life mother idea repulsive spent life trying survive know feel joy body failing memory getting worse worse day pain good point alive joy love card dysphoria transgender medically transition family disown strange think hrt happy thing transspaces looking trans people transitioned feel hopeless want badly know happen live independently thought checking hospital abused 14 transitioned terrified event repeating transgender people treated professional work community know know love alive stand life precious good thing enjoy happy loved life perfect set happy insult people saw given seen poverty disease death abuse help survived happy people drowned afloat unable joy right nose wrong terrifying thought like suicidal miserable faking case die right better happiness lie medium society	Non-suicide	Suicide

Table 11 Confusion matrix of model for CEASEv2.0 dataset

Predicted Values	Actual Values	
	Positive (1)	Negative (0)
Positive (1)	537	93
Negative (0)	154	203

Impact of dataset on model performance The properties of the datasets used in our study considerably influence model's design and performance. The datasets differ concerning the kind of text (written notes versus social media posts), the platform of origin (Reddit versus suicide notes), and the classification task (binary versus multi-class).

For instance, the SuicideDetection and CEASEv2.0 datasets from Reddit involve informal conversational language, which is typical of social media platforms. These datasets are noisy and contain a wide variety of language terms, demanding substantial preprocessing phases such as slang normalization and context interpretation. In contrast, the SWMH dataset, consisting of suicide notes, includes more formal and structured language. Each dataset requires unique preprocessing strategies to manage its specific characteristics.

Additionally, the binary classification tasks (suicidal vs. non-suicidal) in the SuicideDetection and CEASEv2.0 datasets are fundamentally different from the multiclass classification task in the SWMH dataset (depression, suicide watch, anxiety, off-my-chest, and bipolar). The multiclass classification requires more complicated models capable of distinguishing between tiny changes in text indicative of distinct mental health states. We used transformer-based models such as BERT for the SWMH dataset to utilize their superior ability in handling context and capturing semantic nuances.

Error analysis

In this section, we analyze the erroneous cases, i.e., those user comments that have been misclassified by our models. These examples are listed in Table 10. We reported shorter examples taken from the SuicideDetection dataset., while the majority of sentences in this dataset are longer. The reasons for this misclassification could be different. Firstly, for the shorter sentences, some features that affect the output could mislead the result. In the first comment of Table 10, the words "tell", "story" and "wanted" mislead the model to misclassify the comment as suicidal, because suicidal texts usually contain these words. Therefore, some features have a greater effect on shorter sentences than they have on longer ones. In the third comment, the author has used some nonsense words that make the comment look like slang. In the last comment, the repetition of some words, whether in the shorter or longer text, could also lead to misclassification. Suicidal ideation detection requires a semantic understanding of text, which highlights the importance of structured text, far from slang/informal language.

Some texts contain sarcastic sentences, in which the actual meaning of the sentence is different from what appears superficially. This issue makes it difficult to classify these types of comments. For example, in the sixth example of Table 10, the statement 'pretty good life' is positive, and the frequency of the word 'happy' is 4. Therefore, the model is unlikely to detect suicidal ideation in this comment.

Another challenge is quotations because quotations imply the feelings or thoughts of another person, while the model tries to recognize the feelings of the author. This issue will end up in misclassification if the quoted text and the authors' text are semantically different. This issue will be more challenging if the quotations are not written between quotation marks.

Impact of Model Biases and Errors on Downstream Tasks The presence of biases in AI-based models can significantly impact their effectiveness and fairness, particularly in applications such as suicidal ideation detection. When a model is trained on a specific dataset, it may inherit biases related to demographic representation, language usage, and cultural contexts. These biases can lead to skewed predictions that can disproportionately affect certain populations. For example, language nuances and idiomatic expressions unique to specific communities may not be adequately captured, resulting in higher rates of false negatives or false positives, which impacts the performance of the model on downstream tasks such as content moderation and clinical screening.

In clinical screening, biases and inaccuracies could have serious consequences. False negative cases mean missing the opportunity to intervene in cases of true risk, potentially leading to adverse outcomes for the individuals concerned. Although false positives are less critical, they can still cause anxiety by focusing on people who are not at risk.

To mitigate these issues, it is vital to test the models on datasets that reflect different demographics and linguistic styles. Furthermore, including human specialists (human-in-the-loop approach [50]) in critical decision-making processes ensures that professional judgment complements automated systems rather than being replaced by them (Table 11).

Limitations

Detection of suicidal ideation through social media has several limitations. Self-reported data from social media may not accurately represent an individual's true feelings or experiences. Even individuals may write sarcastic sentences that can be entirely misunderstood.

Besides, there are ethical implications for data accessibility, i.e., an individual's privacy should be preserved while using his/her posts. Additionally, the demographic characteristics of the users, such as age, gender, and socioeconomic status, may not be easily inferred from social media data due to privacy issues.

Moreover, the temporal nature of social media posts can make it difficult to detect the suicide risk emergently, as suicide posts may be written only hours or days before the attempt.

Note that two proposed models out of three in this paper perform binary classification, while the third one performs a multi-class classification. Another challenge is the degree of suicidal ideation due to the risk of binary classification, where the model has to assign a 0 or 1 label to people (suicidal or non-suicidal) without taking into account the risk level.

Despite these challenges, the proposed models still play a crucial role in identifying patterns and trends in suicidal ideation and depression on social media, which can be improved by the expertise and knowledge to guarantee precise and prompt interventions.

Practical and Ethical issues. The proposed model can be integrated into mental health care systems to provide early warning indications of suicide ideation, potentially saving lives by allowing timely intervention. Concretely, mental health practitioners can utilize these tools to monitor social media platforms for psychological distress signals and reach out to those exhibiting signs of suicidal thinking.

However, deploying these models has some challenges. One of them is the accuracy and reliability of the model across different populations and contexts. Variability in language use, cultural differences, and the use of slang or sarcasm can negatively affect the model's performance. The false negative cases are more important and dangerous than false positive ones. As mentioned earlier, it is not so dangerous to estimate a normal person as having suicide ideation but estimating someone at high risk of suicide as normal may result in losing a person's life. Another challenge is maintaining users' privacy and complying with legal regulations, such as GDPR [48], which requires careful handling of personal data and explicit consent from users. Anonymizing the user data is an important aspect of such research works. Transparency in such AI-based estimation systems is also essential to develop trust among users and stakeholders. The datasets we used in this research work have no personal information such as names or user IDs-user reviews are anonymous-otherwise, we would have to get permission from ethical commissions.

Conclusion & future work

In this study, we developed three suicidal ideation detection models using classic machine and deep learning algorithms including Transformer applied to three different datasets, two from the Reddit platform and one from suicidal notes. Our motivation was the urgent need to detect and prevent self-harm among people with mental disorders, particularly those who express suicidal thoughts in their online posts.

The proposed models outperform the state-of-the-art in two out of three datasets. Our obtained performance in the Suicide Detection and CEASE-v2.0 datasets in terms of F1 score are 0.97 (0.9742 accuracy) and 0.75, outperforming the state-of-the-art accuracy of 0.95 [5] and 0.74 [16] F1 scores, respectively. However, our model did not achieve state-of-the-art performance on the SWMH dataset, with an F1 score of 0.68 ranking second after the 0.72 F1 score of [24].

Depending on the class labels of the dataset, Our models accomplished binary and multi-class classification.

The current research approves the potential of NLP, ML, and DL methods for suicide prevention. These techniques can be specifically used by psychologists to detect those social media users who have the potential to self-harm. leading to even more refined and broad language representations.

The current study focused solely on Reddit data, but future work will consider other social media platforms such as Twitter and Facebook. In future work, we will also consider a combination of textual, visual, video, and audio data. Besides, we will also consider users' personal characteristics such as age, gender, location, and hobbies.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availibility statement The Datasets that are used in this work can be accessed via the following links. SuicideDetection Dataset: https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch. CEASEv2.0 Dataset: This dataset can be accessed by sending a request to the authors of its paper: [16]. SWMH Dataset can be accessed on https://zenodo.org/records/6476179 after sending a request to its creator.

Declarations

Conflict of interest There are no conflict of interest to disclose for the current research work.

References

1. (2021) Suicide. https://www.who.int/news-room/fact-sheets/detail/suicide
2. Abdulsalam, A., & Alhothali, A. (2022). Suicidal ideation detection on social media: A review of machine learning methods. arXiv preprint arXiv:2201.10515
3. Agarap, A.F. (2018). Deep learning using rectified linear units (relu). CoRR abs/1803.08375. http://arxiv.org/abs/1803.08375, arXiv:1803.08375
4. Agbe(JCharis) EJesse   (2023) Neattext. https://jcharis.github.io/neattext/
5. Aldhyani, T.H.H., Alsubari, S.N., & Alshebami, A.S., et al. (2022). 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). https://doi.org/10.3390/ijerph191912635, https://www.mdpi.com/1660-4601/19/19/12635
6. Ansari, L., Ji, S., Qian, C., et al. (2023). Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on Computational Social Systems, 10*, 211–219. https://doi.org/10.1109/TCSS.2022.3154442

7. Benton, A., Mitchell, M., & Hovy, D. (2017). Multi-task learning for mental health using social media text. https://doi.org/10.48550/arXiv.1712.03538, arXiv:1712.03538
8. Bojanowski, P., Grave, E., & Joulin, A., et al. (2016). Enriching word vectors with subword information. CoRR abs/1607.04606. , arXiv:1607.04606
9. Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. F. Soulié & J. Hérault (Eds.), *Neurocomputing* (pp. 227–236). Berlin Heidelberg, Berlin, Heidelberg: Springer.
10. Cohen, M. R., & Maunsell, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience, 12*(12), 1594–1600.
11. Coppersmith, G., Dredze, M., & Harman, C., et al. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Association for Computational Linguistics, Denver, Colorado, pp 31–39, https://doi.org/10.3115/v1/W15-1204, https://aclanthology.org/W15-1204
12. Coppersmith, G., Leary, R., Crutchley, P., et al. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights,10*, 1178222618792860. https://doi.org/10.1177/1178222618792860, pMID: 30158822
13. Devlin, J., Chang, M., & Lee, K., et al. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. , arXiv:1810.04805
14. Dozat, T. (2016). Incorporating nesterov momentum into adam
15. Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2020). CEASE, a corpus of emotion annotated suicide notes in English. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 1618–1626, https://aclanthology.org/2020.lrec-1.201
16. Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2021). A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation,14*,. https://doi.org/10.1007/s12559-021-09828-7
17. Grant, R., Kucher, D., León, A., et al. (2018). Automatic extraction of informal topics from online suicidal ideation. *BMC Bioinformatics,19*,. https://doi.org/10.1186/s12859-018-2197-z
18. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm networks. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., IEEE, pp 2047–2052
19. Guntuku, S. C., Yaden, D. B., Kern, M. L., et al. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences,18*, 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005, https://www.sciencedirect.com/science/article/pii/S2352154617300384, big data in the behavioural sciences
20. Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. CoRR abs/2104.05314., arXiv:2104.05314
21. Ji, S., Yu, C., Sf, F., et al. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity, 2018*, 1–10. https://doi.org/10.1155/2018/6157249
22. Ji, S., Li, X., & Huang, Z., et al. (2020). Suicidal ideation and mental disorder detection with attentive relation networks. CoRR abs/2004.07601. , arXiv:2004.07601
23. Ji, S., Li, X., Huang, Z., et al. (2021). Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications, 34*(13), 10309–10319. https://doi.org/10.1007/s00521-021-06208-y
24. Ji, S., Zhang, T., Ansari, L., et al. (2022). MentalBERT: Publicly available pretrained language models for mental healthcare. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 7184–7190, https://aclanthology.org/2022.lrec-1.778
25. Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. International Conference on Learning Representations
26. Komati, N. (2021). Suicide and depression detection. www.kaggle.com/datasets/nikhileswarkomati/suicide-watch
27. Kondrak, G. (2005). N-gram similarity and distance. In: International symposium on string processing and information retrieval, Springer, pp 115–126

28. Lai, S., Liu, K., He, S., et al. (2016). How to generate a good word embedding. *IEEE Intelligent Systems, 31*(6), 5–14. https://doi.org/10.1109/MIS.2016.45
29. Loper, E, & Bird, S. (2002). Nltk: the natural language toolkit. CoRR cs.CL/0205028. https://doi.org/10.3115/1118108.1118117
30. Losada, D. E., & Crestani, F., et al. (2016). A test collection for research on depression and language use. In N. Fuhr, P. Quaresma, & T. Gonçalves (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 28–39). Cham: Springer International Publishing.
31. Mikolov, T., & Chen, K., & Corrado, G., et al. (2013). Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR 2013
32. Moulahi, B., Azé, J., & Bringay, S. (2017). Dare to care: A context-aware framework to track suicidal ideation on social media. pp 346–353, https://doi.org/10.1007/978-3-319-68786-5_28
33. Namin, A., Leboeuf, K., & Muscedere, R., et al. (2009). Efficient hardware implementation of the hyperbolic tangent sigmoid function. pp 2117 – 2120,https://doi.org/10.1109/ISCAS.2009.5118213
34. Narayan, S. (1997). The generalized sigmoid activation function: Competitive supervised learning. *Information Sciences,99*(1), 69–82. https://doi.org/10.1016/S0020-0255(96)00200-9, https://www.sciencedirect.com/science/article/pii/S0020025596002009
35. Pennebaker, J., Francis, M., & Booth, R. (1999). Linguistic inquiry and word count (liwc)
36. Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
37. Pirina, I., Çöltekin. (2018). Identifying depression on reddit: The effect of training data. pp 9–12, https://doi.org/10.18653/v1/W18-5903
38. Qader, W., M. Ameen, M., & Ahmed, B. (2019). An overview of bag of words;importance, implementation, applications, and challenges. pp 200–204, https://doi.org/10.1109/IEC47844.2019.8950616
39. Ramírez-Cifuentes, D., Freire, A., & Baeza-Yates, R., et al. (2020). Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis. Journal of Medical Internet Research 22
40. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, https://arxiv.org/abs/1908.10084
41. Richardson, L. (2007). Beautiful soup documentation
42. Sawhney, R., Joshi, H., & Gandhi, S., et al. (2020). A time-aware transformer based model for suicide ideation detection on social media. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 7685–7697, https://doi.org/10.18653/v1/2020.emnlp-main.619, https://aclanthology.org/2020.emnlp-main.619
43. Shah, F.M., Haque, F., & Un Nur, R., et al. (2020). A hybridized feature extraction approach to suicidal ideation detection from social media post. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp 985–988, https://doi.org/10.1109/TENSYMP50017.2020.9230733
44. Shing, H.C., Nair, S., & Zirikly, A., et al. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. Association for Computational Linguistics, New Orleans, LA, pp 25–36, https://doi.org/10.18653/v1/W18-0603, https://aclanthology.org/W18-0603
45. Sinha, P., Mishra, R., & Sawhney, R., et al. (2019). #suicidal - a multipronged approach to identify and explore suicidal ideation in twitter. pp 941–950, https://doi.org/10.1145/3357384.3358060
46. Tadesse, M.M., Lin, H., & Xu, B., et al. (2020). Detection of suicide ideation in social media forums using deep learning. Algorithms 13(1). https://doi.org/10.3390/a13010007, https://www.mdpi.com/1999-4893/13/1/7
47. Vaswani, A., Shazeer, N., & Parmar, N., et al. (2017). Attention is all you need. CoRR abs/1706.03762. http://arxiv.org/abs/1706.03762, arXiv:1706.03762
48. Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed, Cham: Springer International Publishing 10(3152676):10–5555
49. Wolf, T., Debut, L., & Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Online, pp 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6

50. Wu, X., Xiao, L., Sun, Y., et al. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems,135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014, https://www.sciencedirect.com/science/article/pii/S0167739X22001790
51. Xu, S., & E S, Xiang Y,. (2020). Enhanced attentive convolutional neural networks for sentence pair modeling. *Expert Systems with Applications,151*, 113384.
52. Zhang, L., & Moldovan, D. (2019). Multi-task learning for semantic relatedness and textual entailment. *Journal of Software Engineering and Applications, 12*, 199–214. https://doi.org/10.4236/jsea.2019.126012

Authors and Affiliations

Özay Ezerceli[1] · Rahim Dehkharghani[2]

✉ Rahim Dehkharghani
rahim.dehkharghani@khas.edu.tr
https://scholar.google.com/citations?user=Fjn6GQgAAAAJ&hl=en&oi=ao

Özay Ezerceli
ozay.ezerceli@isikun.edu.tr
https://scholar.google.com/citations?hl=en&user=Z6tvnkEAAAAJ

[1] Computer Engineering, Isik University, Mesrutiyet, Sile 34980, Istanbul, Turkey

[2] Management Information Systems and Computer Engineering Departments, Kadir Has University, Cibali campus, Fatih 34083, Istanbul, Turkey