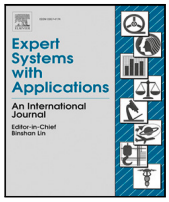
Expert Systems With Applications 255 (2024) 124855

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A self-attention TCN-based model for suicidal ideation detection from social media posts

Seyedeh Leili Mirtaheri *, Sergio Greco, Reza Shahbazian

Department of Informatics, Modeling, Electronics and System Engineering (DIMES), University of Calabria, Italy

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Suicidal ideation
Social networks
Natural language processing

ABSTRACT

Early suicidal ideation detection has long been regarded as an important task that can benefit both society and individuals. In this regard, it has been shown that, very frequently, the first symptoms of this problem can be identified by analyzing the contents shared on social media. Machine learning classification models have proven promising in capturing behavioral and textual features from posts shared on social media. This study proposes a novel machine-learning model to detect the risk of suicide from social media posts, employing both natural language processing and state-of-the-art deep learning techniques. We propose an ensemble LSTM-TCN model that benefits from a self-attention mechanism to detect suicidal ideation among users of two well-known social networks, Twitter (X) and Reddit. Furthermore, we present a comprehensive analysis of the data, examining the suicidal posts both statistically and semantically, which can provide rich knowledge about suicidal ideation. Our proposed model (AL-BTCN) outperforms the compared state-of-the-art models, resulting in over 94% accuracy, recall, and F1-score. Researchers, mental health specialists, and social media service providers can all benefit from the findings of this study.

1. Introduction

Suicide has long been considered a major challenge in the realm of mental health due to its seriousness and unique characteristics. According to the World Health Organization (WHO), more than 703,000 individuals worldwide suffer from suicidal ideation by committing suicide annually. Suicide has been considered one of the leading causes of death, particularly among the younger generation and people in deprived countries, which accounts for around 77% of global suicide cases. Early detection of suicidal ideation could prevent several suicide cases (Pigoni et al., 2024). Therefore, it is important to promptly identify the possibility of suicide ideation. The fact that many people – roughly 1 in every 6 – have to wait to receive psychological care (Bruen, Wall, Haines-Delmont, & Perkins, 2020) supports the significance of early detection of suicidal conditions.

Background. Numerous studies have been carried out on the early detection of suicidal conditions. Long-term predictors have been identified by working with hospitalized clinical patients (Balbuena et al., 2022), while possible short-term predictors, for example, on emotional agony, social problems, and sleep problems have also been addressed in the literature (Allen, Nelson, Brent, & Auerbach, 2019).

With advancements in technology, we are witnessing the increase in automated and data-driven approaches for suicidal ideation detection. An important challenge in data-driven methods is acquiring adequate and comprehensive data. Social networking platforms can play a critical role in facilitating data collection since they have transformed conventional communication methods (Islam et al., 2023). These platforms enable users to express their feelings and share their thoughts. Social networking platforms have become an inseparable part of our daily lives and can be considered one of the main sources of information, as their openness can provide an opportunity for people to share their thoughts, ideas, and opinions (Brailovskaia, Margraf, Schillack, & Köllner, 2019). As a result, monitoring them can be an efficient approach for early suicide detection (Kailasam & Samuels, 2015). Moreover, there has been an increasing willingness to express suicidal thoughts on social media and online communication forums. This characteristic of social media platforms, in addition to the possibility of being anonymous, has provided an opportunity for people to express even their negative feelings and emotions openly (Mbarek, Jamoussi, & Hamadou, 2022). Consequently, the data on social media can provide a rich understanding of individuals' emotions and suicide risk. This can be regarded as a complementary source of knowledge for clinical purposes (Ji, Yu, Fung, Pan, & Long, 2018).

Data extracted from social media can be analyzed using classical machine learning algorithms. Traditional neural networks, specifically convolutional neural networks (CNNs) and long short-term memory

* Corresponding author.
E-mail addresses: leili.mirtaheri@dimes.unical.it (S.L. Mirtaheri), greco@dimes.unical.it (S. Greco), reza.shahbazian@unical.it (R. Shahbazian).

https://doi.org/10.1016/j.eswa.2024.124855
Received 14 February 2024; Received in revised form 12 June 2024; Accepted 20 July 2024
Available online 23 July 2024

(LSTM), have been largely used in the context of mental health issues. However, these methods suffer from not fully extracting the features of the textual contents. Therefore, the methods that rely only on combining CNN and LSTM, are not capable of effectively capturing deep text features (Kour & Gupta, 2022b). Furthermore, these approaches suffer from scalability difficulties, making them less effective for large-scale analysis. The majority of existing models are memory intensive, which limits their practical application. Also, traditional RNN-based methods suffer from vanishing gradient issues, which have a severe impact on training and performance (Agarwal, Mahajan, Shrotriya, & Shekhawat, 2024). This means that there is still a need for more precise models to effectively and accurately detect suicidality among social media content. This need, motivated us to do this research. We focus on textual information because it is the most popular type of shared content on social media.

Problem statement. The aim of this paper is the early identification of suicidal ideation based on the analysis of social network content. Our goal is to propose a method that resolves the existing issues in the literature while improving both the performance and scalability.

Proposed solution. We propose an advanced machine learning model that aims to distinguish suicidal social media posts from normal (non-suicidal) ones. In our proposed model, we combine state-of-the-art deep learning, namely temporal convolutional networks (TCN), self-attention, and LSTM techniques, to improve both performance and scalability compared to existing methods. By using LSTM, we resolve the vanishing gradient issue introduced by conventional RNNs. We enhance the performance by adding Bi-directional TCN layers to capture long-range temporal patterns in both forward and backward directions. This combination enables our proposed model to extract information more reliably. Longer memory and improved context understanding are guaranteed by implementing the bi-directional TCN layers. In addition, we provide an in-depth analysis of the datasets used in our experimentation. We compare the proposed method with state-of-the-art techniques (Ghosh et al., 2023; Renjith, Abraham, Jyothi, Chandran, & Thomson, 2022) and baseline ML classifications. The results confirm that the proposed model can extract the features more comprehensively and in more general terms. Our model requires less memory, is scalable, and avoids common exploding and vanishing gradient issues. Based on the inherent nature of TCNs (Bai, Kolter, & Koltun, 2018), our proposed model can show longer memory in comparison with well-known RNN-based methods.

Contributions. The main contributions of the paper are as follows:

- We propose a novel self-attention-based LSTM bidirectional TCN (Bi-TCN) model to distinguish suicidal ideation from typical, non-suicidal social media posts. Our proposed model detects suicidal posts with higher accuracy thanks to the use of an ensemble model built by using LSTM, self-attention, and the bidirectional architecture of TCN.
- We identify prominent and distinguishable features of suicidal thoughts in comparison to the typical, non-suicidal ones, present in the textual contents shared by users on social networks (Twitter and Reddit). We use these features to better predict suicidal ideation in social media posts. Our analysis includes both lexical and semantic aspects, which helps extract relevant features from the data, leading to more accurate detection.
- To evaluate the proposed framework, we perform an extensive comparison with the baseline machine learning methods, including Random Forest, Decision Tree, Gradient Boost, and Logistic Regression, as well as the state-of-the-art hybrid attention-based methods presented in Ghosh et al. (2023) and Renjith et al. (2022).

Organization. The remainder of this paper is as follows. Section 2 explores the recent literature on suicidal ideation detection and provides a thorough overview of the present datasets and features most commonly used in the literature. Section 3 is devoted to the proposed method, beginning with an overview of the model and then presenting its specific features. Section 4 presents an in-depth analysis of the model setup considered, the experimental results, and their comparison with state-of-the-art approaches. Finally, Section 5 provides our conclusions, faced challenges and proposes potential future directions for this research.

2. Related works

There are a considerable number of studies devoted to suicidality detection in the literature. Generally, the related works can be classified into two categories, namely clinical or computer-based studies. As shown in Fig. 1, clinical studies mainly aim to consider three objectives consisting of (i) identifying potential lifetime predictors, (ii) suicidality risk assessment, and (iii) feasibility studies to highlight the use of modern technology-based instruments in gathering the required data, such as electronic health records (EHR) (Boudreaux et al., 2021; Gupta, Bhatia, & Kumar, 2021; Sanderson, Bulloch, Wang, Williamson, & Patten, 2020; Thiruvalluru, Gaur, Thirunarayan, Sheth, & Pathak, 2021), questionnaires or smartphone applications (Ayhan, Üstün, & Ergüze, 2020; Balbuena et al., 2022; Sels et al., 2021) or even wearables and sensors (Kleiman et al., 2019).

As illustrated in Fig. 2, computer-based suicide ideation detection studies usually utilize textual (Gaur et al., 2019; Tadesse, Lin, Xu, & Yang, 2019), image (Cao et al., 2019; Lekkas, Klein, & Jacobson, 2021), video (Nguyen, Kavuri, & Lee, 2019), and audio contents. They aim to detect suicidal ideation by using deep learning models, in addition to transfer learning, and time-aware-based techniques (Adarsh, Kumar, Lavanya, & Gangadharan, 2023; Cai, Wang, Ye, Jin, & Gao, 2023; Dheeraj & Ramakrishnudu, 2021). Deep learning and hybrid approaches have been claimed to be highly successful in detecting suicidality with noticeable accuracy (Amanat et al., 2022; Haque, Islam, Islam, & Ahsan, 2022; Islam, Qusar, & Islam, 2021; Kour & Gupta, 2022a; Renjith et al., 2022). Moreover, fuzzy interference systems and network-based classification approaches such as attention networks (AN) have been popular in recent years (Mbarek et al., 2022; Vashishtha & Susan, 2020).

Lekkas et al. utilize the ensemble learning model on social media data to detect suicidal ideation (Lekkas et al., 2021). In their proposed method, the main idea is to combine the conventional classifiers. Abdulsalam, Alhothali, and Al-Ghamdi (2024) studied suicidal ideation in Arabic tweets, exploring the potential of different deep learning models. They show the effectiveness of machine learning models in detecting suicidality among social networking posts. In another work, Ghanadian, Nejadgholi, and Al Osman (2024) focused on the data scarcity issue and proposed a synthetic data strategy using large language models (LLMs). Their experiments with suicidal ideation detection across different classifiers show that accuracy may improve by using synthetic data.

Comparing the performance of machine learning models, the authors in Ao, Lai, Lu, and Mo (2021) show that the RNN-based models, particularly the LSTM, can outperform traditional models. The application of natural language processing techniques to automatically detect suicidality among English tweets has been proposed in Metzler, Baginski, Niederkrotenthaler, and Garcia (2022). The authors indicate that deep learning models are promising in the field of suicidal thought detection compared with conventional word frequency classifiers (Metzler et al., 2022). Kodati and Tene (2024) review the potential of deep learning models in suicidality risk detection among social media content. Similarly, the authors in Boonyarat, Liew, and Chang (2024) and Kancharapu and Ayyagari (2024) explore enhanced BERT and GAN-infused deep learning models in suicidal risk assessment through social media posts and illustrate the practicality of more advanced deep learning models. In addition, a time-aware transfer-based model, combining linguistic models with historical contexts, to identify suicidal

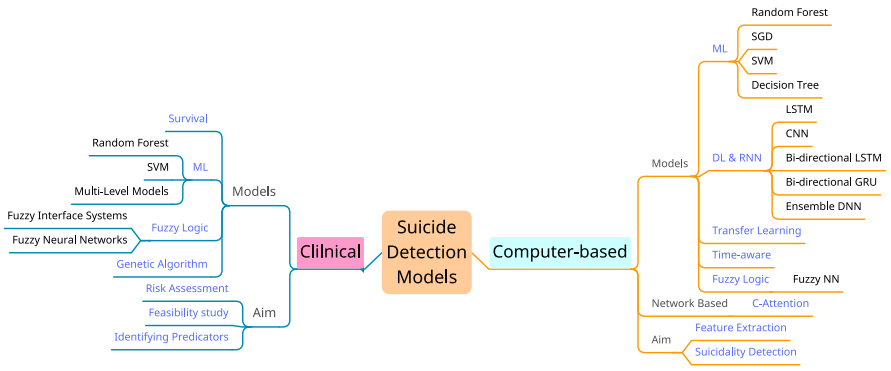


Fig. 1. Overall view of the techniques that related literature that uses machine learning for suicide detection models.

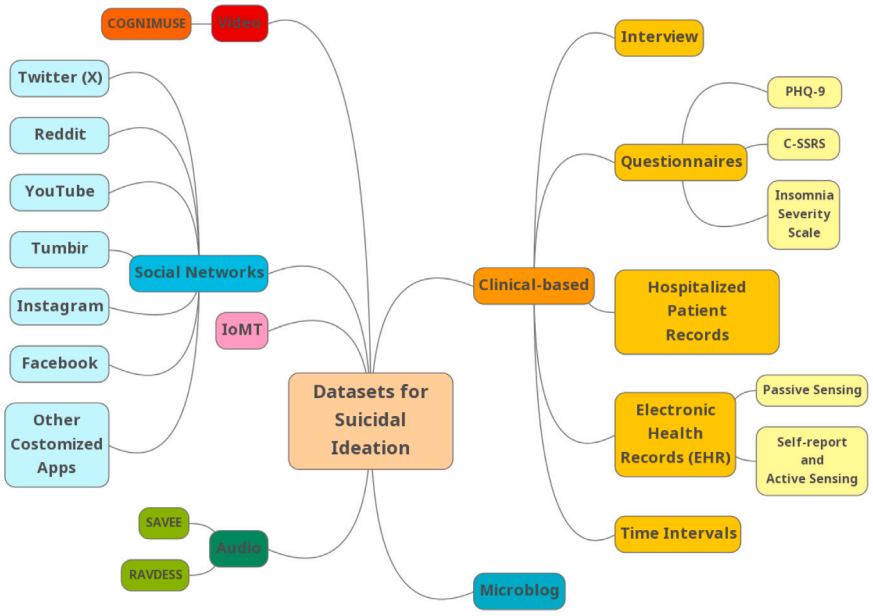


Fig. 2. The overall view of different datasets used in the literature for suicide ideation.

intent on Twitter is proposed by Sawhney, Joshi, Gandhi, and Shah (2020).

By drawing attention to important parts of the data, feature engineering can also help suicidal ideation detection models (Amanat et al., 2022; Lekkas et al., 2021). As an example, Bhattacharya et al. proposed to use a *bag-of-words* feature extraction technique as a prediction tool on the Twitter dataset (Bhattacharya, Shahina, et al., 2021). In another study, Ghosh et al. (2023) proposed an architecture using the combination of the long attention mechanism, LSTM, and CNN layers to detect the depressive posts.

Based on the results of the above-mentioned research, it is critical to identify the indicators and factors with more importance and to place greater emphasis on these certain features and aspects. State-of-the-art hybrid techniques that combine LSTM and CNN, while incorporating attention have been proposed to address these issues (Ghosh et al., 2023; Renjith et al., 2022). These techniques provide improvements in capturing data features and emphasizing relevant indicators compared with conventional deep learning methods.

A summary of the related works is presented in Table 1, where we categorize the related works based on their implemented techniques and the datasets. As can be seen in Table 1, despite the contributions of existing studies, there is still room for further development. The existing methods are not able to fully extract the features of the textual contents. This motivated us to propose a new and advanced architecture for suicidal ideation detection that is scalable, can better extract the features and guarantees efficacy, as discussed in Section 3.

3. Proposed method

The aim of this paper is the early identification of suicidal ideation based on the analysis of social network content. Thus, this section describes the proposed framework for detecting the posts that might contain information about suicidal intent.

The main contribution and novelty of the paper is the proposal of a neural network architecture for detecting suicidal intents that combines long short-term memory (LSTM) and bidirectional temporal convolutional neural network (Bi-TCN). LSTM is a type of RNN that is mainly used to learn, process, and classify sequential data. Common LSTM applications include sentiment analysis, language modeling, speech recognition, and video analysis. It is also aimed at dealing with the vanishing gradient problem present in traditional RNNs. Bidirectional TCN is a neural network architecture designed for sequence modeling and has recently been proposed to overcome some limitations of the temporal convolutional network (TCN) (Bahdanau, Cho, & Bengio, 2014; Zhang, Ma, & Liu, 2022).

The present study proposes an integrated deep learning model called AL-BTCN, combining LSTM and Bi-TCN. The proposed framework is shown in Fig. 3, which comprises the following modules: *Pre-Processing*, *Word embedding*, and *Classifier*. First, the input data, consisting of

Table 1
Categorization of the related works based on their implemented techniques, features and datasets.

Ref.	Features	Techniques/methods	Data source/Dataset	Category
Balbuena et al. (2022)	Predictive factors for suicide	Survival models, - machine learning models	Hospital/Clinic visits	Clinical
Ayhan et al. (2020)	Reliability and validity - of modern tools	Fuzzy logic	Questionnaires	Clinical
Sels et al. (2021)	Connection between predictors and suicidal outcomes	Machine learning models	Applications	Clinical
Kleiman et al. (2019)	Behavioral, psychological, and physiological factors	Passive and active sensing	Wearable Sensors	Clinical
Boudreaux et al. (2021) and Gupta et al. (2021)	Detecting suicidality - among users	Electric health record analysis	EHRs	Clinical
Thiruvalluru et al. (2021)	Psychological risk factors for suicide	Clustering, semantic similarity	EHRs, Reddit	Clinical
Islam et al. (2021)	Suicidal ideation detection	Machine learning	Facebook	Clinical
Sanderson et al. (2020)	AI techniques for suicidal ideation	XGB, RNN	EHR	Clinical
Amanat et al. (2022) and Renjith et al. (2022)	Suicidality detection	Deep learning, hybrid ensemble learning	Muti	Clinical
Adarsh et al. (2023) and Cai et al. (2023)	Detection of suicidal ideation	Deep learning, transfer learning	Social Media	Computer-based
Dheeraj and Ramakrishnudu (2021)	Suicidal ideation detection	Time-aware techniques	Social Media	Computer-based
Vashishtha and Susan (2020)	Fuzzy interference systems	Neuro-fuzzy networks, - attention network models	Social Media	Computer-based
Cao et al. (2019) and Wang et al. (2021)	Behavioral, emotional, - location-based features	Machine learning, NLP	Social Networks	Computer-based
Amanat et al. (2022) and Lekkas et al. (2021)	Feature engineering, - suicidality detection	Ensemble learning models	Social Media	Computer-based
Bhattacharya et al. (2021)	Suicide - prediction tools	Bag-of-Words	Twitter	Computer-based
Ghosh et al. (2023)	Deep learning models	Long attention mechanism, LSTM, CNN	Social Media	Computer-based
Ao et al. (2021)	NLP techniques for tweets	RNN, LSTM	Social Media	Computer-based
Metzler et al. (2022)	Suicidal thought detection	Deep learning models	Twitter	Computer-based
Sawhney et al. (2020)	Linguistic models - with historical context	Time-aware transfer-based model	Twitter	Computer-based
Abdulsalam et al. (2024)	Suicidality detection - in Arabic content	Machine learning models	Social Media	Computer-based
Ghanadian et al. (2024)	Synthetic data generation	Large language models	UMD Dataset	Computer-based
Kodati and Tene (2024)	Emotion mining	Deep learning Dynamic masking-based transformer	Social Media, Suicidal Notes	Computer-based
Kancharapu and Ayyagari (2024)	Genetic optimization	GAN, deep learning models	Social Media	Computer-based
Boonyarat et al. (2024)	Suicidality detection in Thai social media	Bert, deep learning models	Social Media	Computer-based

textual contents, gathered from social networking websites (posted by users), goes through a pre-processing procedure to remove possible existing noise from the raw data before being passed to the subsequent phase. Next, the output of the *Pre-Processing* module goes as input to the *Word Embedding* module, whose outputs, consisting of n-grams, are given as input to the *Classifier*. Ultimately, the output of the classifier states whether a post contains relevant suicidal intent.

The classifier takes the vectors resulting from the embedding model and classifies them based on the presence of information indicating a suicide risk or not. As illustrated in Fig. 3(C), the first layer consists of an LSTM network that captures the long-term dependencies among the words, and the output is fed into two sequential Bi-TCN layers to capture the long-range temporal patterns over the texts. Next, the pipeline features an attention layer, which helps highlight prominent information by assigning weights to words, followed by two average and maximum pooling layers, each of which processes the data individually, and then the obtained results are concatenated to form a final vector before the classification process comes into play.

Finally, two dense layers are provided to classify the concatenated vector into a single output, indicating the relevance of the suicidality intention contained in the post, using the rectified linear unit (ReLU) and the sigmoid activation functions, respectively. It should be noted that two dropout layers were also included in the model, one after the embedding layer and one before the final dense layer, to avoid possible overfitting of the model over the train data. The detailed architecture of our proposed classifier is presented in Fig. 4, whereas a more detailed description of the modules is next provided.

3.1. Pre-processing

As presented in Fig. 3, the pre-processing procedure filters data to preclude any possible noise and misleading content existing in the collected raw data, entering the embedding model and, consequently, interfering with the learning process. Such unwanted content would prevent the model from correctly learning crucial information and, therefore, from correctly classifying the analyzed texts. To do so, the textual content collected from social networks is usually cleaned and filtered from punctuation marks and newline characters, Uniform Resource Locators (URLs), hashtags, mention characters, and things as such. Furthermore, all textual content was made lowercase.

Therefore, all posts were filtered to ensure no stop word would be included in the texts. Subsequently, the posts were tokenized using the *Natural Language Toolkit* (Bird, 2006) so that every sentence would be transformed into a list of tokens. The cleaned and filtered dataset would be fed into the embedding layer to generate the required vectors.

3.2. Word embedding

To capture the contextual, semantic, and syntactic features of each word in a document, word embedding techniques provide a representation of a word in the form of a real-valued vector. This representation

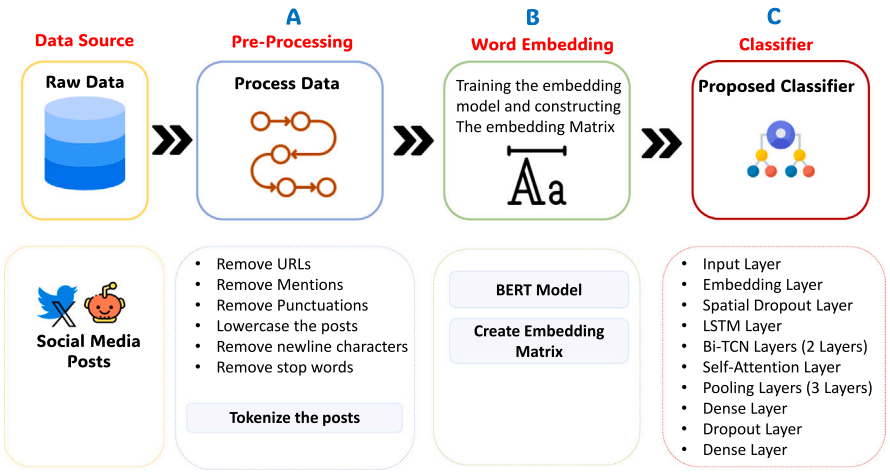


Fig. 3. The general architecture of the proposed model (AL-BTCN) for suicidal ideation detection based on social network posts.

ensures that words with closer vectors have higher similarity, indicating their relatedness in meaning and structure

We employ the BERT embedding model (Alsentzer et al., 2019) to produce a context-informed embedding for each word. BERT embedding can lead to more effective capturing of details in texts. However, commonly used word embeddings create the same embedding vector for a word in a sentence, regardless of the differences in the context, which can result in the loss of prominent information. Analyzing social network posts, the meanings associated with the same word can heavily depend on the context in which it is used. As a result, it is essential to capture the subtle and sometimes sophisticated language employed acutely. In this respect, the BERT embedding approach used in this work can recognize the signals and changes in the text, going beyond simple keyword recognition, which enables it to equip the model better to grasp signs of suicidal ideation.

3.3. Classifier

Here, we present the components of the proposed multi-layer classifier, as shown in Fig. 4. The proposed classifier comprises several data processing layers. Concisely, the model begins with receiving the data and preparing the embedding vectors. Then, the embedding layer transforms integer-encoded words into dense vectors representing their semantics. To ensure that the model does not over-rely on any single feature of the word vectors, a spatial dropout layer with a 0.1 ratio is also applied during training, which randomly deactivates entire embedding dimensions for all words in a sequence. Next, the LSTM and two bidirectional TCN layers process the data to extract relevant information, and the self-attention layer and pooling layers are performed on the vectors whose prominent values are highlighted by the attention layer. Concatenating the two vectors received from pooling layers, two fully connected layers with a dropout layer in between are performed on the data to provide the final decision about the data samples. In continuation, we present a more detailed description of the main submodules.

3.3.1. LSTM

The LSTM networks benefit from tackling the vanishing gradient problem, which can be a serious obstacle considering the performance of RNN networks (Brownlee, 2017). LSTM uses three gates, namely the forget gate ($f_t \in [0,1]$), the input gate (i_t), and the output gate (o_t). These gates determine how much data should be forwarded and what should be discarded. Through these gates, the hyperbolic tangent, ($\tanh(.) \in [-1,1]$) and sigmoid functions ($\sigma(.) \in [0,1]$) are applied to manage and update the cell state.

As can be seen in Fig. 5, considering (t, c_t) as the time step and the cell state at the time t, the network will receive the input vector x_t along with the hidden states $h_{(t-1)}$ and the cell state $c_{(t-1)}$ from the previous time step $t-1$. The inputs, outputs, and calculations performed in an LSTM network are given in the next Eqs. (1)–(6).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3}$$

$$\hat{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

$$C_t = f_t \odot C_{(t-1)} + i_t \odot \hat{C}_t \tag{5}$$

$$h_t = o_t \odot \tanh(C_t) \tag{6}$$

where the initial values of the LSTM's memory cell state are $c_0 = 0$ and $h_0 = 0$, the operator $\odot$ denotes the Hadamard product (element-wise product), and o_t denotes the output for the current time step, along with c_t and h_t which are generated for the next time step. The subscript t indexes the time step. In these equations, (f_t, i_t, o_t) refers to the gates forget, input, and output, respectively, and (u_t, c_t, h_t) depicts the hyperbolic tangent layer, the memory cell, and the hidden states at time step t, respectively. Moreover, W represents the weight matrix, and b depicts the bias vector used during the calculations in each gate. In the proposed method, an LSTM layer including 100 units is employed, which receives the embedded input vectors from the spatial dropout layer and creates an output vector consisting of a sequence of values.

3.3.2. Bi-TCN

A temporal convolutional network (TCN) framework processes sequential data, combining the notions of casual convolutions and dilation (Bai et al., 2018). TCN models benefit from two main characteristics: (i) the casualty of the convolutional layers that result in preventing information leakage from later time steps to former ones, and (ii) its architecture that enables this model to map an input sequence of any length to its corresponding output of the same length (Bai et al., 2018). Eq. (7), where term $x_{(s-d\cdot i)}$ denotes the data point in the sequence that is dilated by $(d \cdot i)$ steps from the current position, k denotes the filter size, and d is the dilation factor, shows how the dilated convolution

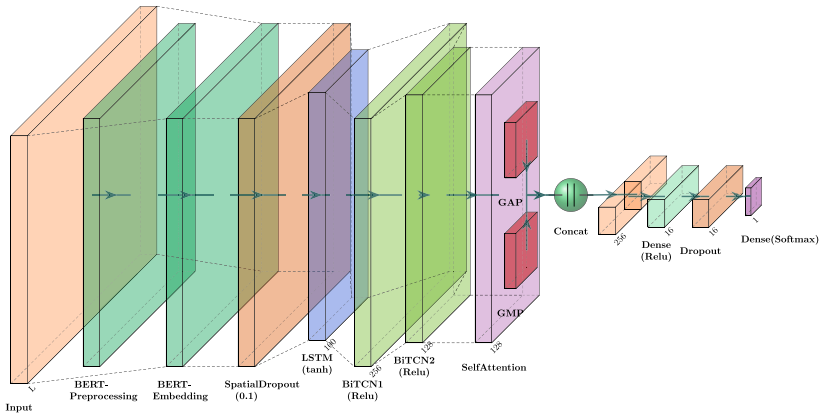


Fig. 4. The architecture of the proposed AL-BTCN classifier. In a TCN, a residual block is used between the input and the output layers.

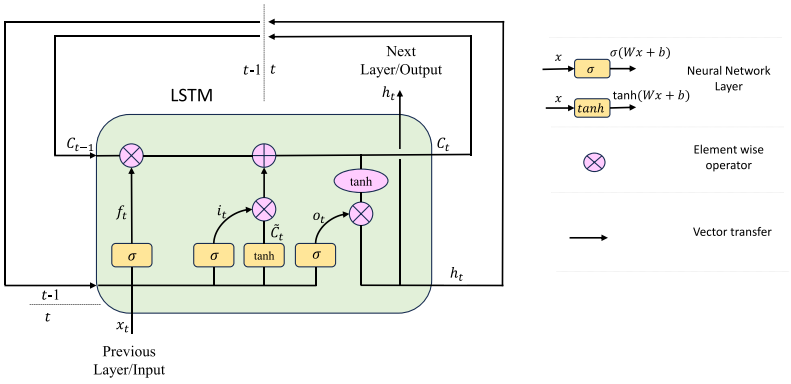


Fig. 5. The architecture of LSTM layer used in the proposed architecture.

operation will be done on a one-dimensional sequence of input, using a filter $f : [0, k-1] \rightarrow R$.

$$F(s) = \sum_{(i=0)}^{(k-1)} f(i) \cdot x_{(s-d \cdot i)} \tag{7}$$

When two of these convolution layers are stacked together, a residual block is constructed. In addition, provided that the lengths of the input and output data differ from each other, we add a one-dimensional convolution to create a residual block to produce the appropriate vector with the same length as the output. The residual block is employed to facilitate the training of deep models and accelerate convergence. Therefore, a TCN will be constructed by stacking any required number of these residual blocks according to the problem at hand. A temporal convolutional network is made up of a dilated casual convolution and the TCN-corresponding residual block, which can be seen in Fig. 6.

Despite the diverse applications of TCN models, they have been proven to outperform the LSTM models in natural language processing context (Zhang et al., 2022). Bidirectional layers enable the processing of the input sequences in both forward and backward manner. This bidirectional information flow is commonly used in tasks requiring context from past and future inputs. For instance in text classification, it helps to capture dependencies and make informed predictions.

The proposed AL-BTCN model utilizes two consequent hidden layers of Bidirectional TCN (Bi-TCN) to fully capture the contextual information from the embedded vectors obtained from applying the embedding techniques and representing various words in posts (Zuo et al., 2020). According to our experiments, using two TCN layers in the bidirectional form leads to higher performance, in particular regarding the recall metric, resulting in a more accurate classification of the suicidal posts. Moreover, by integrating the bidirectional concept with TCN layers, the model can benefit from capturing dependencies in both directions of the input sequence, enhancing the overall performance of TCN layers, which leads to improved suicidal ideation detection.

3.3.3. Attention layer

The attention mechanism improves the accuracy of machine translation tasks (Bahdanau et al., 2014). In the proposed AL-BTCN model, a self-attention layer is added after the TCN layers to extract and highlight the crucial information existing in the input posts given to the model as inputs. The attention mechanism is particularly beneficial for suicide detection in social networks, as it allows the model to focus on specific words in a post, ensuring that crucial signals are not overlooked.

3.3.4. Dropout

The dropout layer is a technique used to prevent the model from overfitting. This layer ensures that our model does not rely too heavily on any particular neuron, promoting a more generalized model. In addition to the dropout layers, AL-BTCN also makes use of a spatial dropout layer, which is a variation of the former one but is specifically used in convolutional networks. The spatial dropout layer is applied after the embedding layer, and a dropout layer is added before the final Dense layer, each with a 0.1 drop-rate.

3.4. Integration of the layers

Here, we discuss the integration of the layers in our proposed architecture. As shown in Fig. 4, the proposed model consists of several data processing layers. First, we utilize the BERT embedding layer to generate numerical word-embedding vectors for the input text contents while capturing their semantics in the texts. BERT is a pre-trained language model, trained on an extensive amount of data, which allows the

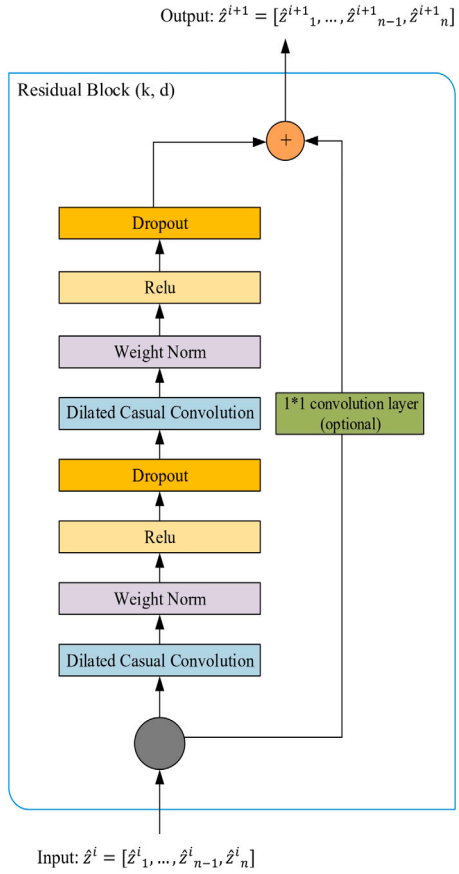


Fig. 6. The architecture of the residual block in the TCN layers used in the proposed architecture.

model to develop a profound understanding of language and makes it effective for natural language processing applications. The multi-layer bidirectional transformers architecture of BERT, enables the model to better comprehend the full contextual meaning of each word. This lead to enhanced understanding of the overall text.

LSTM layers are able to process sequential data efficiently and selectively remember, update, and retrieve information over extended sequences. These networks can capture patterns and dependencies in the text content, along with effectively adapting to dynamic input sequences. Furthermore, a model architecture utilizing LSTM can handle the vanishing gradient problem due to its embedded gating mechanism. On the other hand, using dropout layers prevents the model from overly relying on any single feature of the word vectors and mitigates the potential overfitting problem.

The data is processed by the LSTM and two bidirectional TCN layers to extract relevant information, followed by the self-attention layer and pooling layers, with the attention layer highlighting the most prominent values. The attention mechanism examines the correspondence of a sequence of key–value pairs and assesses the likeness between the associated keys and values within the pairs. This assessment yields a weight, enhancing the precision of text classification by distributing varying significance to the textual content through the allocation of distinct weights. On the other hand, TCN layers use causal convolutions to prevent information leakage.

The use of bidirectional layers in deep learning allow processing of input sequences in both forward and backward directions, enabling the network to capture context from past and future inputs. As a result, the proposed AL-BTCN model uses two consecutive bidirectional TCN layers, following the LSTM layer, to fully capture contextual information from embedded vectors. Our experiment results (see Section 4), indicate that the usage of the bidirectional TCN architecture, along with the LSTM layer, enhances the model's ability to capture dependencies in text contents. This leads to higher performance and results in more accurate classification of suicidal posts. We utilize two types of pooling layers: max and global average pooling. Pooling layers are used to reduce the noise and the likelihood of overfitting. The max pooling layer reduces computational operations by decreasing the spatial dimensions in the feature map. Additionally, the global average pooling better aligns with the convolution structure, enforcing connections between feature maps and classes of data.

Table 2
Experimental setup of the proposed AL-BTCN model.

Main layers		
Layer	Parameters	Setting
BERT pre-process		
BERT encoder	Model	Small bert en_uncased_L4_H512_A8
LSTM	Units, activation function	100, tanh
Bidirectional TCN #1	Units, activation function, dilations	128, ReLU, [1, 2, 4]
Bidirectional TCN #2	Units, activation function, dilations	64, ReLU, [1, 2, 4]
Remaining layers		
Spatial dropout	Dropout rate	0.1
Dense	Units	16, ReLU
Dropout	Dropout rate	0.1
Dense	Units	1, Sigmoid

4. Experimental results

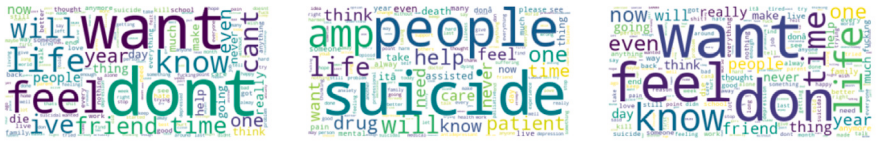
To evaluate the proposed model, we trained AL-BTCN on data collected from Twitter and Reddit and performed extensive experiments to compare its effectiveness with state-of-the-art techniques. As baseline methods, we consider Random Forest, Decision Tree, Gradient Boost, and Logistic Regression. We also evaluate the performance of deep learning-based methods, including LSTM-CNN, LSTM, and CNN. As state-of-the-art methods, we consider two ensemble learning models: Ghosh et al. (2023) and Renjith et al. (2022). An LSTM-Attention-CNN model, reporting an accuracy of 90.3% and an F1-score of 92.6% on the Reddit dataset, is proposed in Zirikly, Resnik, Uzuner, and Hollingshead (2019). In Ghosh et al. (2023) the ensemble model is comprised of LSTM and CNN layers, accompanied by a *Luong Attention* layer, and reports an accuracy of 94.3% on both Bangala and English Twitter datasets. We implement these models based on their settings presented in the paper, using "word2vec" for the model in Renjith et al. (2022) and "fasttext embedding" methods for the model in Ghosh et al. (2023). The source code for our experiments is publicly available at the Github web site.[1]

We use a GeForce RTX 2080 GPU and 45 GB of RAM, Python 3.7 as the programming language, and the Keras open-source library (Tensorflow), in addition to the Natural Language Toolkit. The TCN-based layers are implemented using the Keras Temporal convolutional network library. We train the proposed model on 30 epochs with a batch size of 64, utilizing the stochastic gradient descent (SGD) optimizer and "binary categorical cross-entropy" loss function. The number of units, other related settings, and the model architecture are illustrated in Table 2.

[1] https://github.com/ShahbazianR/TCN-based-Model-for-Suicidal-Ideation-Detection-from-Social-Media-Posts.git.

Table 3
Statistical information of the datasets used in our experiments.

Dataset	Total posts	Suicidal posts	Non-suicidal posts
Twitter_1 (TwitterDS1, 2018)	9119	3998	5121
Twitter_2 (TwitterDS2, 2020)	17 142	3754	13 388
Reddit_SNS (RedditDS1, 2021)	232 074	116 037	116 037



(a) Twitter_1 (b) Twitter_2 (c) Reddit_SNS

Fig. 7. The word clouds associated with the Twitter_1, Twitter_2, and Reddit_SNS datasets.

4.1. Dataset

To evaluate the performance of the proposed AL-BTCN, three publicly available datasets are used, including textual content collected from two popular social networks: Twitter and Reddit. Table 3 illustrates the statistics of the datasets used in the following experiments. To preserve user privacy, the personal information has been either replaced with a unique ID. Datasets include binary labeled data on suicidal and non-suicidal posts. Reddit_SNS is created by more than 200,000 posts from this social networking site, including more than 100,000 posts for each class of suicidal or non-suicidal content shared on the subreddits. The suicidal posts often account for a small percentage of total content on social networks. In our experiments, the Twitter_2 dataset is unbalanced. To balance the data and preventing unaccurate training, we add suicidal-labeled posts provided in ScuidalTweets (2018) limiting the characters to a maximum number of 200 tokens to satisfy the constraints on Twitter_2 dataset. In the pre-processing, we filter this raw data to make sure that all stop word are removed from the text. We further transform the sentences into a list of tokens.

4.2. Data analysis

We inspect the datasets from various aspects, including the cloud of words, n-grams, topics, polarity, and semantic analysis.

4.2.1. Word cloud

As illustrated in Fig. 7, the majority of suicidal posts in Twitter_1 (Fig. 7(a)) and the Reddit dataset (Fig. 7(c)) have mentioned the words “want”, “feel”, and “know”, mainly in their negative forms such as “don’t know”, “can’t” or “don’t want”, which can indicate the implied message of depression or suicidal ideation by the reluctance used in these messages. In the Twitter_2 dataset (Fig. 7(b)), the most relevant word is “suicide”, but very often they also refer to other people by using the general term “people”. Moreover, terms like “life”, “die”, “think”, “need”, “help”, “will”, and “time” have been frequently mentioned in all datasets, although they are used less than other formerly mentioned words.

4.2.2. N-grams

We extracted lists of bi-grams and tri-grams to identify which n-grams are mostly used in suicidal ideation posts.

(A) **Twitter_1**. The most frequent bi-grams are “don’ t know”, “feel like” and “don’ t want”, each accounted for 1061, 982 and 923 times, respectively. The tri-grams, “don’t want live”, “don’t even know”, “don’t know anymore”, “don’t want die”, “want die want”, “can’t take anymore”, “don’t feel like”, “don’t see point”, “don’t know much”, “want live anymore” are the top-10 tri-gram terms mentioned in this datasets of tweets, repeated for 85 down to 34 times, in order.

(B) **Twitter_2**. The bi-grams such as “assisted dying”, “mental health”, “palliative care”, “assisted suicide”, “mental-health bipolar” and “feel like” were repeated in the tweets up to 138 times. Moreover, tri-gram terms such as “bipolar anxiety ocd”, “suicide-prevention fear anxiety”, “attempt suicide problem”, “investigation life ends”, “care desperately need”, “struggling mental health” and “palliative care desperately”, which occur more than 30 times, indicate the presence of mental health problems, such as bipolar disorder, depression, and suicide anxiety in phrases.

(C) **Reddit_SNS**. The bi-tram “feel like” appeared to be the most frequently used, with more than 1662 occurrences. Furthermore, “plz plz”, “want die”, “don’t know” and “don’t want”, each posted more than 340 times, in addition to “want kill”, “need help”, “wish could” and “want end”, which were also mentioned frequently (up to 240 times). The most frequent tri-grams are “plz plz plz”, “help need help”, “pain sad pain”, “things get better”, “really want die”, “feel like shit”, “need someone talk”, “life worth living”, “feel like one” and “want live anymore”.

The above-mentioned bi-grams and tri-grams show that people with suicidal ideation tend to reflect their depressive and negative feelings and emotions through their posts, either implicitly or explicitly. These phrases usually include the negation forms of “want” or “take”, meaning that suicidal people are typically reluctant to continue their suffering.

Furthermore, suicidal posts may also contain rather positive terms such as “get better”, “high school”, and “best friends”, although their frequencies are much lower than those of negative feelings and contents denoting mental problems. As a result, it is important to grasp the importance of monitoring changes in mood in sequential time intervals and considering the whole context of their posts rather than focusing on single or multiple terms.

4.2.3. Frequent topics

Table 4 reports the four main topics related to suicides for each dataset. These topics reveal not only the presence of suicidal ideation but also reflect depressive thoughts. Many users show a tendency to talk to someone and attempt suicide, which can be found in sentences such as “need someone talk” and “attempt suicide”. It is also worth mentioning that topics such as “get better” could indicate either progress or wishing to get better, which in either case can be an early sign of suicide detection.

4.2.4. Polarity and subjectivity

Sentiment analysis can assist in better capturing various features present in the textual content. To this end, we use a pre-trained library named “TextBlob”, which is provided in the Natural Language Toolkit. Polarity is a score that shows the sentiment of the sentence and lies in

Table 4
The four top frequent topics of suicide ideation identified during the experiments on Twitter and Reddit datasets.

	Twitter_1	Twitter_2	Reddit_SNS
Topic #1	really want die	bipolar disorder	feel like
Topic #2	depressed go sorry	attempt suicide, suicide problem resolved	need someone talk
Topic #3	want fucking die,	investigation life ends, life ends	want die
Topic #4	Don' t want, don't know, feel like	attempting suicide, huge family	Don' t know, I' m tired

Table 5
Polarity and Subjectivity results gained from the experiments on Twitter and Reddit datasets. The parameter P stands for polarity and S indicates the subjectivity.

Dataset	$P < 0$	$(P = 0)$	$(P > 0)$	$(S > 0.5)$
Twitter_1	44.886%	9.17%	45.93%	52.73%
Twitter_2	33.50%	13.27%	53.23%	50.91%
Reddit_SNS	47.93%	4.43%	49.09%	58.94%

the range $[-1, 1]$, where negative values indicate negative sentiment, whereas positive values refer to positive sentiment. Subjectivity is a score in the range of $[0, 1]$ to indicate how much a sentence can be considered either a personal opinion or a factual statement. The low values of the subjectivity score refer to no personal opinion (i.e., factual statement), and as the score goes higher, the degree of personal opinion increases.

According to Table 5, almost similar percentages of suicidal posts in all datasets are present, as indicated by negative $(P < 0)$ or positive $(P > 0)$ sentiment, making up approximately half the suicidal social media posts. In contrast, the posts showing a neutral sentiment $(P = 0)$ account for less than 15% of the whole posts with suicidal ideation in both Twitter and Reddit datasets. Also, $S > 0.5$ suggests that the majority of posts are opinionated in nature.

4.2.5. Semantic analysis

Table 6 illustrates the results obtained from applying a pre-trained 3-labeled semantic categorization to the experimented datasets. It should be noted that labels for posts, in terms of being negative, neutral, and positive, are assigned based on the base of the highest score. For example, the tweet "depression I want to die" scored 0.768 as negative, 0.101 as neutral, and 0.131 as positive, therefore, the label assigned to this post is negative. However, the compound measure computes the aggregate sentiment of the entire text. It takes into account all the neutral, positive, and negative sentiments and provides a single score by averaging over the individual scores. As presented in Table 6, the majority of posts in each dataset are labeled as neutral (around 90%), while considering the compound metric, the majority of the posts could be categorized as positive or negative (more than 95%).

4.3. Comparison results

Here, we report the performance results of the proposed AL-BTCN and its comparison with baselines and state-of-the-art techniques.

4.3.1. Baseline methods

We have implemented Random Forest, Decision Tree, Gradient Boost, and Logistic Regression as baseline methods. As shown in Table 7, Random Forest could result in better performance with accuracy and recall values of over 83% on all datasets. While Gradient Boost and Logistic Regression could detect suicidal posts with similar performance on Twitter datasets; their performance results significantly dwindled on Reddit_SNS. The experimental results also demonstrate that the proposed AL-BTCN noticeably outperforms traditional classification models in terms of accurately detecting suicidality in textual content shared in social networks, resulting in around 94% recall and F1-score, and approximately 95% accuracy, 95% precision on the experimented datasets.

4.3.2. Deep learning methods

Table 8 presents the experiment results for LSTM, CNN, and LSTM-CNN compared to our proposed AL-BTCN model. The settings employed for each of the considered deep learning approaches is described in details in the followings:

The LSTM-CNN is designed as an ensemble model consisting of one layer LSTM with 100 units, a CNN layer with 3 filters, a kernel size of 8, and a ReLU activation function. A dropout layer with a rate of 0.5 is used after the initial Input layer. A flattened layer is before the final dense layer with size 1 and a Sigmoid activation function, which will do the final classification. On the other hand, The LSTM model considered in the experiments consists of one LSTM layer with 100 units followed by the input layer, accompanied by a dropout with a 0.05 rate, and a flattened layer before and after it, respectively. Finally, a dense layer with the same setting is considered for the final classification. Finally, the CNN model begins with an input layer, followed by a dropout layer with the same rate as above. Then, a one-dimensional convolution layer with a filter size of 3 and kernel size of 8 is considered for the convolution layer, with the Leaky ReLU layer as its activation function. The model ends with a flattened and dense layer similar to the above-mentioned settings to classify the data by providing the probability or likelihood of a post being suicidal.

As can be seen, the LSTM model shows the highest accuracy and F1-score among the deep learning models on the Twitter_1 dataset with 85% accuracy and 82% F1-score, respectively (Table 8). LSTM-CNN and CNN each experienced higher evaluation findings in recall and precision, accordingly, on this dataset. Regarding the other datasets, all three models produced similar results within a range of 50% to 60%. However, LSTM-CNN and LSTM resulted in higher accuracy, precision, recall, and F1-score in comparison to the other deep learning-based baselines on Reddit_SNS. Moreover, all the compared deep learning models failed to accurately detect suicidality in the data collected from Twitter_2 (Table 8), accounting for a performance of approximately 50% in all evaluation metrics. The results of this experiment show that the proposed AL-BTCN model shows the highest accuracy, precision, recall, and F1-score on all datasets, outperforming the deep learning baselines considered in this section with 92% accuracy on Twitter-based datasets and more than 94% accuracy on Reddit_SNS (Table 8).

4.3.3. Ensemble learning models

Table 9 summarizes the comparison results of the proposed AL-BTCN with the state-of-the-art ensemble learning models LSTM-Attention-CNN and BiLSTM-CNN, presented in Ghosh et al. (2023) and Renjith et al. (2022), denoted as LSTM-Attention-CNN and BiLSTM-CNN, respectively. In comparison to the considered benchmark models, on all datasets, the proposed AL-BTCN outperforms the experimented models in all the metrics. Considering the recall metric, that is frequently emphasized in the context of suicidal ideation detection, the average improvement is of the 3%.

4.4. Contribution of the AL-BTCN's components

In this section, we present a comparative analysis of the contribution of the proposed AL-BTCN model's components. To evaluate the effect of the considered components and the utilized BERT embedding, we remove one layer each time, and the resulting model is trained and tested. Since the model cannot be implemented without embedding, we use Word2Vec (W2V) embeddings instead of BERT in our evaluations. The derived frameworks, obtained by deleting a single component, are resumed in Table 10, where the models are named on the basis of the remaining components. The description of each model is as follows:

Table 6
The results of semantic categorization on the evaluated datasets. In the first section, we report the sentiment based on the highest score, while using the compound metric, we report the average score.

Dataset	Negative	Neutral	Positive	Compound < 0	Compound = 0	Compound > 0
Twitter_1	4.27%	95.22%	0.50%	73.31%	1.67%	25.01%
Twitter_2	1.89%	97.46%	0.64%	59.49%	4.94%	35.57%
Reddit_SNS	0.08%	99.04%	0.78%	71.50%	1.49%	27.01%

Table 7
Comparison of AL-BTCN with standard baseline models (Random Forest, Decision Tree, Gradient Boost, and Logistic Regression) in terms of accuracy, precision, recall, and F1-score. The results have been rounded to the second decimal digit.

Datasets	Models	Accuracy	Precision	Recall	F1-Score
Twitter_1	Random Forest	0.88	0.88	0.88	0.88
	Decision Tree	0.86	0.86	0.86	0.86
	Gradient Boost	0.88	0.88	0.88	0.87
	Logistic Regression	0.86	0.87	0.86	0.85
	AL-BTCN	**0.94**	**0.94**	**0.91**	**0.92**
Twitter_2	Random Forest	0.84	0.84	0.84	0.84
	Decision Tree	0.64	0.64	0.64	0.64
	Gradient Boost	0.69	0.69	0.69	0.69
	Logistic Regression	0.56	0.56	0.56	0.54
	AL-BTCN	**0.92**	**0.91**	**0.94**	**0.92**
Reddit_SNS	Random Forest	0.83	0.83	0.83	0.83
	Decision Tree	0.65	0.65	0.65	0.65
	Gradient Boost	0.69	0.69	0.69	0.69
	Logistic Regression	0.56	0.56	0.56	0.54
	AL-BTCN	**0.95**	**0.95**	**0.94**	**0.95**

Table 8
Comparison of the LSTM, CNN, and LSTM-CNN with the proposed AL-BTCN model in terms of accuracy, recall, precision, and F1-score.

Datasets	Models	Accuracy	Precision	Recall	F1-Score
Twitter_1	CNN	0.84	0.94	0.67	0.78
	LSTM	0.85	0.89	0.76	0.82
	LSTM-CNN	0.85	0.86	0.78	0.82
	AL-BTCN	**0.94**	**0.94**	**0.91**	**0.92**
Twitter_2	CNN	0.59	0.58	0.58	0.58
	LSTM	0.58	0.58	0.55	0.57
	LSTM-CNN	0.58	0.59	0.54	0.60
	AL-BTCN	**0.92**	**0.91**	**0.94**	**0.92**
Reddit_SNS	CNN	0.56	0.56	0.57	0.56
	LSTM	0.67	0.68	0.63	0.65
	LSTM-CNN	0.68	0.69	0.66	0.67
	AL-BTCN	**0.95**	**0.95**	**0.94**	**0.95**

Table 9
Comparison of proposed AL-BTCN with state-of-the-art LSTM-Attention-CNN (Renjith et al., 2022) and BiLSTM-CNN (Ghosh et al., 2023) in terms of accuracy, precision, recall and F1-score.

Dataset	Model	Accuracy	Precision	Recall	F1-score
Twitter_1	LSTM-Attention-CNN	0.92	0.93	0.90	0.91
	BiLSTM-CNN	0.92	0.91	0.90	0.90
	AL-BTCN	**0.94**	**0.94**	**0.91**	**0.92**
Twitter_2	LSTM-Attention-CNN	0.89	0.90	0.88	0.89
	BiLSTM-CNN	0.90	0.90	0.90	0.90
	AL-BTCN	**0.92**	**0.91**	**0.94**	**0.92**
Reddit_SNS	LSTM-Attention-CNN	0.91	0.92	0.90	0.91
	BiLSTM-CNN	0.91	0.89	0.93	0.91
	AL-BTCN	**0.95**	**0.95**	**0.94**	**0.95**

- A-2BTCN is obtained by removing the LSTM layer. Thus, it consists of one attention layer and two layers of bidirectional TCN. Moreover, the pooling, dropout, concatenation, and dense layers are intact.
- AL-1BTCN is obtained by removing the second TCN layer. The obtained model includes the defined attention mechanism, the LSTM layer, one bidirectional TCN layer, pooling, dropout, concatenation, and dense layers.

Table 10
Definition of the model names to evaluate the effect of utilized layers on the performance of the proposed AL-BTCN.

Model	Removed component
A-2BTCN	LSTM layer
AL-1BTCN	1 BiTCN layer
AL	TCN layers
L-2BTCN	Attention layer

- AL is obtained by removing both TCN layers. The resulting model includes the attention mechanism, the LSTM layer, pooling, dropout, concatenation, and dense layers.
- L-2BTCN is obtained by only removing the attention layer, whereas all the other layers remain intact.

Fig. 8 shows the accuracy, precision, recall, and F1-score achieved after deleting the layers of the proposed model on the Twitter_2 dataset. It demonstrates that deleting layers from the proposed model reduces performance, particularly in terms of recall. The results also reveal that TCN layers in our proposed AL-BTCN play a significant role, as removing them decreases performance by around 60% in terms of recall.

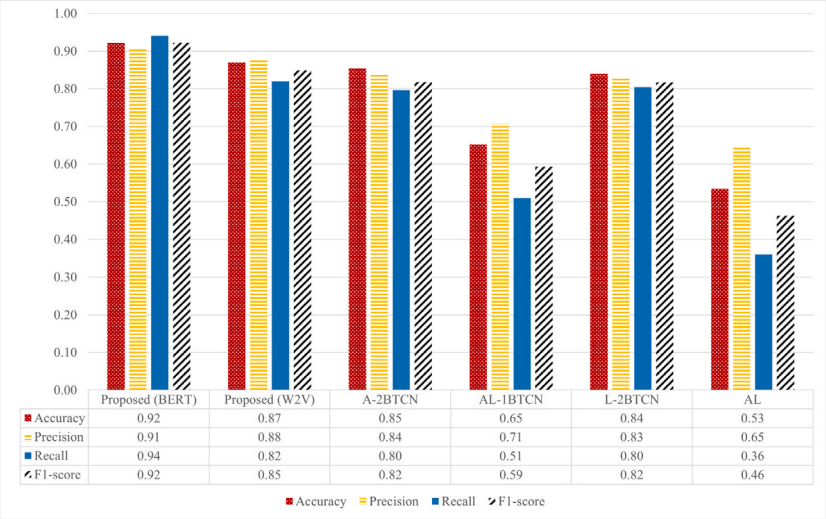
5. Conclusions and future directions

In this paper, we examined the use of deep learning techniques to detect suicidal ideation among social media posts. We proposed a novel self-attention-based ensemble learning model that benefits from LSTM and bi-directional TCN layers to detect the risk of suicidality among posts collected from social media and distinguish such posts from typical, non-suicidal ones. This model aims to assist healthcare professionals by early detection of suicidal ideation among users on social networking platforms, and ultimately intends to help them reduce the rate of suicides committed by individuals in society. According to the experimental results gained from the conducted evaluations, the proposed model can accurately detect suicidal ideation among users, outperforming both rather traditional baselines, machine learning, and state-of-the-art deep learning models. With an accuracy of 95% and over 92% on Reddit and Twitter datasets, respectively, a recall score of 95%, and an F1-score of over 92%, the proposed model's performance was witnessed to be more accurate and higher than other baseline techniques. The proposed AL-BTCN model is able to accurately detect suicidal thoughts from social media posts and holds great potential for suicide prevention initiatives across various sectors. It allows mental health professionals to identify at-risk individuals, improve the efficacy of suicide prevention measures, and enhance user trust and loyalty. Businesses and social networking sites can monitor suicidal content, incorporating safety measures and alarm systems. Mental health support networks can use this model to enhance outreach and response plans, ensuring timely assistance and a decrease in suicides.

5.1. Challenges

Here, we state the challenges faced during conducting this study, and present implemented approaches to address such issues.

The first challenge was related to the scarcity of adequate and comprehensive suicidality data. Trusted data typically could be collected through clinical procedures which are not usually publicly available



	Proposed (BERT)	Proposed (W2V)	A-2BTCN	AL-1BTCN	L-2BTCN	AL
Accuracy	0.92	0.87	0.85	0.65	0.84	0.53
Precision	0.91	0.88	0.84	0.71	0.83	0.65
Recall	0.94	0.82	0.80	0.51	0.80	0.36
F1-score	0.92	0.85	0.82	0.59	0.82	0.46

Fig. 8. The contribution of layers in the proposed AL-BTCN model. We use a stochastic gradient optimizer, the Bert and Word2Vector(W2V) embeddings, and the evaluations are performed on the Twitter_2 dataset. We use W2V as the default embedding, and the definitions of the model names are presented in Table 10.

due to privacy concerns. To mitigate this issue, we focused on social networking platforms. Considering such approach can be beneficial for researchers, psychologists, and can also assist social networking companies to monitor the mental health state of their users. This might help them to develop new strategies (for instance emergency system connections or artificial intelligence chatbots) to take prompt actions when necessary.

Social media datasets are typically unbalanced. This is because suicidal posts accounting for a minor percentage of the total contents on social networks. In our study, the Twitter_2 dataset was unbalanced that could lead to difficulties in developing robust and accurate suicidality detection model. To deal with this issue, we added suicidal notes to balance the Twitter_2 dataset. We used the posts from other social networks and limited the content to a maximum number of 200 tokens to be aligned with the Twitter dataset.

Choosing the exact embedding models, as well as processing layers can be difficult, requiring several experiments (trial and error). We have conducted numerous experiments with various embedding methods and model processing layers. Ultimately, considering our evaluations, we came up with the proposed layer integration. This proposed architecture resulted in the highest rate of performance, in terms of accuracy and error metrics.

5.2. Future works

In terms of future works, we aim to experiment with different combinations of learning algorithms to explore the potentially effective ensemble models, in addition to examining different attention approaches. We believe that the study carried out in this paper can further improved by considering different word embedding techniques and feeding various behavioral and social network profile features to the model. We would also like to consider several social media posts in different consequent time intervals to better detect behavior and mental condition changes. This observed changes in one's social media contents can be processed in addition to other related features such as location, application usage time, and social connection. A further potential research direction is to address personalized suicidal detection systems by integrating the explainable learning methods into the proposed architecture.

CRediT authorship contribution statement

Seyedeh Leili Mirtaheri: Conceptualization, Methodology, Software, Resources, Writing – original draft. **Sergio Greco:** Methodology, Validation, Formal analysis, Writing – review & editing. **Reza Shahbazian:** Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The link to data the software code is provided in the paper.

Acknowledgments

All authors approved the version of the manuscript to be published. We acknowledge financial support from PNRR MUR projects FAIR (PE0000013) and SERICS (PE00000014).

References

Abdulsalam, A., Alhothali, A., & Al-Ghamdi, S. (2024). Detecting suicidality in arabic tweets using machine learning and deep learning techniques. *Arabian Journal for Science and Engineering*, 1–14.

Adarsh, V., Kumar, P. A., Lavanya, V., & Gangadharan, G. (2023). Fair and explainable depression detection in social media. *Information Processing & Management*, *60*(1), Article 103168.

Agarwal, H., Mahajan, G., Shrotriya, A., & Shekhawat, D. (2024). Predictive data analysis: Leveraging RNN and LSTM techniques for time series dataset. *Procedia Computer Science*, *235*, 979–989.

Allen, N. B., Nelson, B. W., Brent, D., & Auerbach, R. P. (2019). Short-term prediction of suicidal thoughts and behaviors in adolescents: Can recent developments in technology and computational science provide a breakthrough? *Journal of Affective Disorders*, *250*, 163–169.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., et al. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Amanat, A., Rizwan, M., Javed, A. R., Abdelhaq, M., Alsaqour, R., Pandya, S., et al. (2022). Deep learning for depression detection from textual data. *Electronics*, *11*(5), 676.

Ao, Z., Lai, J., Lu, W., & Mo, H. (2021). Comparisons of LSTM and other machine learning approaches on predicting suicidal tendency regarding to social media posts. In *2021 international conference on computers and automation* (pp. 19–23). IEEE.

Ayhan, F., Üstün, B., & Ergüze, T. T. (2020). The development of a fuzzy logic model-based suicide risk assessment tool. *The Journal of Neurobehavioral Sciences*, *7*(3), 156–163.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Balbuena, L. D., Baetz, M., Sexton, J. A., Harder, D., Feng, C. X., Boctor, K., et al. (2022). Identifying long-term and imminent suicide predictors in a general population and a clinical sample with machine learning. *BMC Psychiatry*, *22*(1), 120.

Bhattacharya, D., Shahina, A., et al. (2021). Early detection of suicidal tendencies from text data using LSTM. In *2021 innovations in power and advanced computing technologies* (pp. 1–5). IEEE.

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions* (pp. 69–72).

Boonyarat, P., Liew, D. J., & Chang, Y.-C. (2024). Leveraging enhanced BERT models for detecting suicidal ideation in thai social media content amidst COVID-19. *Information Processing & Management*, *61*(4), Article 103706.

Boudreaux, E. D., Rundensteiner, E., Liu, F., Wang, B., Larkin, C., Agu, E., et al. (2021). Applying machine learning approaches to suicide prediction using healthcare data: overview and future directions. *Frontiers in Psychiatry*, *12*, Article 707916.

Brailovskaia, J., Margraf, J., Schillack, H., & Köllner, V. (2019). Comparing mental health of facebook users and facebook non-users in an inpatient sample in Germany. *Journal of Affective Disorders*, *259*, 376–381.

Brownlee, J. (2017). *Long short-term memory networks with python: develop sequence prediction models with deep learning*. Machine Learning Mastery.

Bruen, A. J., Wall, A., Haines-Delmont, A., & Perkins, E. (2020). Exploring suicidal ideation using an innovative mobile app-strength within me: the usability and acceptability of setting up a trial involving mobile technology and mental health service users. *JMIR Mental Health*, *7*(9), Article e18407.

Cai, Y., Wang, H., Ye, H., Jin, Y., & Gao, W. (2023). Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, *217*, Article 119538.

Cao, L., Zhang, H., Feng, L., Wei, Z., Wang, X., Li, N., et al. (2019). Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. arXiv preprint arXiv:1910.12038.

Dheeraj, K., & Ramakrishnudu, T. (2021). Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model. *Expert Systems with Applications*, *182*, Article 115265.

Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., et al. (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference* (pp. 514–525).

Ghanadian, H., Nejadgholi, I., & Al Osman, H. (2024). Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEE Access*.

Ghosh, T., Al Banna, M. H., Al Nahian, M. J., Uddin, M. N., Kaiser, M. S., & Mahmud, M. (2023). An attention-based hybrid architecture with explainability for depressive social media text detection in bangla. *Expert Systems with Applications*, *213*, Article 119007.

Gupta, D., Bhatia, M., & Kumar, A. (2021). Real-time mental health analytics using IoMT and social media datasets: research and challenges. In *Proceedings of the international conference on innovative computing & communication*.

Haque, R., Islam, N., Islam, M., & Ahsan, M. M. (2022). A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning. *Technologies*, *10*(3), 57.

Islam, M. R., Qusar, M. S., & Islam, M. S. (2021). Suicide after facebook posts–an unnoticed departure of life in Bangladesh. *Emerging Trends in Drugs, Addictions, and Health*, *1*, Article 100005.

Islam, M. R., Sakib, M. K. H., Prome, S. A., Wang, X., Ulhaq, A., Sanin, C., et al. (2023). Machine learning with explainability for suicide ideation detection from social media data. In *2023 10th international conference on behavioural and social computing* (pp. 1–6). IEEE.

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

Kailasam, V. K., & Samuels, E. (2015). Can social media help mental health practitioners prevent suicides? *Current Psychiatry*, *14*(2), 37–51.

Kancharapu, R., & Ayyagari, S. N. (2024). Suicidal ideation prediction based on social media posts using a GAN-infused deep learning framework with genetic optimization and word embedding fusion. *International Journal of Information Technology*, 1–17.

Kleiman, E., Millner, A. J., Joyce, V. W., Nash, C. C., Buonopane, R. J., Nock, M. K., et al. (2019). Using wearable physiological monitors with suicidal adolescent inpatients: Feasibility and acceptability study. *JMIR mHealth and uHealth*, *7*(9), Article e13725.

Kodati, D., & Tene, R. (2024). Emotion mining for early suicidal threat detection on both social media and suicide notes using context dynamic masking-based transformer with deep learning. *Multimedia Tools and Applications*, 1–24.

Kour, H., & Gupta, M. K. (2022a). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. *Multimedia Tools and Applications*, *81*(17), 23649–23685.

Kour, H., & Gupta, M. K. (2022b). Hybrid LSTM-TCN model for predicting depression using Twitter data. In *2022 17th international conference on control, automation, robotics and vision* (pp. 167–172). http://dx.doi.org/10.1109/ICARCV57592.2022.10004238.

Lekkas, D., Klein, R. J., & Jacobson, N. C. (2021). Predicting acute suicidal ideation on instagram using ensemble machine learning models. *Internet Interventions*, *25*, Article 100424.

Mbarek, A., Jamoussi, S., & Hamadou, A. B. (2022). An across online social networks profile building approach: Application to suicidal ideation detection. *Future Generation Computer Systems*, *133*, 171–183.

Metzler, H., Baginski, H., Niederkrotenthaler, T., & Garcia, D. (2022). Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach. *Journal of Medical Internet Research*, *24*(8), Article e34705.

Nguyen, T.-L., Kavuri, S., & Lee, M. (2019). A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips. *Neural Networks*, *118*, 208–219.

Pigoni, A., Delvecchio, G., Turtulici, N., Madonna, D., Pietrini, P., Cecchetti, L., et al. (2024). Machine learning and the prediction of suicide in psychiatric populations: a systematic review. *Translational Psychiatry*, *14*(1), 140.

RedditDS1 (2021). data retrieved from Kaggle, https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch.

Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). 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.

Sanderson, M., Bulloch, A. G., Wang, J., Williamson, T., & Patten, S. B. (2020). Predicting death by suicide using administrative health care system data: can recurrent neural network, one-dimensional convolutional neural network, and gradient boosted trees models improve prediction performance? *Journal of Affective Disorders*, *264*, 107–114.

Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. (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* (pp. 7685–7697).

ScuidalTweets (2018). data retreived from Github, https://github.com/IE-NITK/TwitterSuicidalAnalysis.

Sels, L., Homan, S., Ries, A., Santhanam, P., Scheerer, H., Colla, M., et al. (2021). SIMON: a digital protocol to monitor and predict suicidal ideation. *Frontiers in Psychiatry*, *12*, Article 554811.

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

Thiruvalluru, R. K., Gaur, M., Thirunarayan, K., Sheth, A., & Pathak, J. (2021). Comparing suicide risk insights derived from clinical and social media data. *AMIA Summits on Translational Science Proceedings*, *2021*, 364.

TwitterDS1 (2018). data retrieved from Github repository, https://github.com/IE-NITK/TwitterSuicidalAnalysis/tree/master/datasets.

TwitterDS2 (2020). data retrieved from Github repository, https://github.com/laxmimerit/twitter-suicidal-intention-dataset.

Vashishtha, S., & Susan, S. (2020). Unsupervised fuzzy inference system for speech emotion recognition using audio and text cues (workshop paper). In *2020 IEEE sixth international conference on multimedia big data* (pp. 394–403). IEEE.

Wang, N., Luo, F., Shivtare, Y., Badal, V. D., Subbalakshmi, K., Chandramouli, R., et al. (2021). Learning models for suicide prediction from social media posts. arXiv preprint arXiv:2105.03315.

Zhang, Y., Ma, Y., & Liu, Y. (2022). Convolution-bidirectional temporal convolutional network for protein secondary structure prediction. *IEEE Access*, *10*, 117469–117476.

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 24–33).

Zuo, Y., Jiang, L., Sun, H., Ma, C., Liang, Y., Nie, S., et al. (2020). Short text classification based on bidirectional TCN and attention mechanism. *Vol. 1693*, In *Journal of physics: conference series*. IOP Publishing, Article 012067.