Progress in Artificial Intelligence (2024) 13:135–147
https://doi.org/10.1007/s13748-024-00326-z

REGULAR PAPER

Advanced deep learning and large language models for suicide ideation detection on social media

Mohammed Qorich[1] · Rajae El Ouazzani[1]

Received: 27 December 2023 / Accepted: 9 June 2024 / Published online: 21 June 2024
© Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Recently, suicide ideations represent a worldwide health concern and pose many anticipation challenges. Actually, the prevalence of expressing self-destructive thoughts especially on forums and social media requires effective monitoring for suicide prevention, and early intervention. Meanwhile, deep learning techniques and Large Language Models (LLMs) have emerged as promising tools in diverse Natural Language Processing (NLP) tasks, including sentiment analysis and text classification. In this paper, we propose a deep learning model incorporating triple models of word embeddings, as well as various fine-tuned LLMs, to identify suicidal thoughts in Reddit posts. In effect, we implemented a Bidirectional Long Short-Term Memory (BiLSTM), and a Convolutional Neural Network (CNN) model to categorize posts associated with non-suicidal and suicidal thoughts. Besides, through the combination of Word2Vec, FastText and GloVe embeddings, our models learn intricate patterns and prevalent nuances in suicide-related language. Furthermore, we employed a merged version of CNN and BiLSTM models, entitled C-BiLSTM, and several LLMs, including pre-trained Bidirectional Encoder Representations from Transformers (BERT) models and a Generative Pre-training Transformer (GPT) model. The analysis of all our proposed models shows that our C-BiLSTM model with triple word embedding and our GPT model got the best performance compared to deep learning and LLMs baseline models, reaching accuracies of 94.5% and 97.69%, respectively. In fact, our best model's capacity to extract meaningful interdependencies among words significantly promotes its classification performance. This analysis contributes to a deeper understanding of the psychological factors and linguistic markers indicative of suicidal thoughts, thereby informing future research and intervention strategies.

Keywords Bidirectional long short-term memory (BiLSTM) · Convolutional neural network (CNN) · Large language models (LLMs) · Mental health · Natural language processing (NLP) · Suicide ideation (SI) · Text classification · Triple word embedding

1 Introduction

Worldwide, Suicide Ideation (SI) is a complex concern and a deep defiance within the realm of public health. Concisely, it refers to considerations, thoughts, or plans related to ending one's own life [1]. Therefore, suicide represents the most extreme manifestation of SI, as it involves harmful self-directed acts in an attempt to end one's life [1]. In fact, suicide ranks as the fifteenth most common source of death contributing to 1.4% of total mortality, as reported by the World Health Organization (WHO) in 2014 [2]. Moreover, about 10.6 million American adults grapple with severe SI annually, according to the Substance Abuse and Mental Health Service Administration (SAMHSA) in 2018 [3], which reflects the common problem of SI affecting individuals across all age groups, genders, and backgrounds. Actually, several factors could be tragic sources of this rampant phenomena including psychological, social, and environmental elements [4]. For instance, social isolation, mental illness, financial difficulties, substance abuse, and past trauma can cause developing suicidal thoughts and behaviors [5–7]. Alternatively, peer victimization and cyberbullying stand out as prevalent and serious factors igniting SI, especially among adolescents [8–10].

✉ Mohammed Qorich
mohamedqorich@gmail.com

Rajae El Ouazzani
elouazzanirajae@gmail.com

[1] IMAGE Laboratory, ISNET Team, School of Technology, Moulay Ismail University of Meknes, Meknes, Morocco

In general, experiencing SI does not necessarily mean behaving on these thoughts. However, it is a critical warning sign of struggling with mental health, that requires intervention and support. Basically, the social stigma surrounding individuals with SI exerts a notably profound influence. In fact, some researchers have demonstrated that the apprehension of social stigma can be an obstacle and discourage individuals from openly discussing their feelings and seeking the required support from others [11, 12]. Nevertheless, social media contribute to the growing prevalence of individuals sharing self-destructive thoughts and experiences [13–15]. Due to the anonymity and relative privacy of some platforms, some individuals are encouraged to disclose their struggles without the fear of immediate personal judgment. Besides, sharing one's personal feelings and experiences within forums dedicated to mental health enhances the sense of support among individuals facing similar challenges. In effect, the individual posts represent a rich source for detecting and analyzing suicidal expressions, sentiments and ideas. Thereby, identifying individuals at risk could help them with the appropriate intervention, such as mental health care, counseling, and crisis management, to promote their overall mental well-being.

On the other hand, Natural Language Processing (NLP) tasks, especially text classification, and sentiment analysis collectively generate a powerful framework for identifying SI in social media contents [16]. NLP techniques allow understanding and processing human language, as well as analyzing a vast amount of text generated data. Along with NLP algorithms, sentiment analysis can determine posts that display signs of despair, distress, hopelessness, or any other indices of potential suicidal thoughts. Besides, text classification involves labeling text data into predefined classes. In the context of suicide prevention, the aim is to classify social media posts into categories such as "suicidal ideation" or "no suicidal ideation".

In this paper, we develop a system to detect suicidal thoughts in Reddit posts. The detection is accomplished through the integration of various deep learning architectures and Large Language Models (LLMS), along with NLP algorithms, especially text classification ones. Effectively, we integrated the Convolutional Neural Network (CNN) and the Bidirectional Long Short-Term Memory (BiLSTM) with various word embedding models. Additionally, we fused these two deep learning models to create a Convolutional BiLSTM (C-BiLSTM) representation, linked to a combined word embedding layer that encompasses three extensively preprocessed embedding models. Moreover, we fine-tuned pre-trained LLMs such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-training Transformer (GPT). Thus, our suggested deep learning model, incorporating triple word embedding, and the GPT model have significantly enhanced the accuracy of identifying suicidal thoughts, leading to effective performance.

The contributions of our paper are:

- A proposed C-BiLSTM model with a triple word embedding technique for SI detection in Reddit posts.
- A comparative analysis of several word embeddings with BiLSTM and CNN for binary text classification.
- A research experiment examining the effects of combining various word embedding pre-trained models, either in pairs or triples, on the classification result.
- A deployment of LLMs for the mental health classification and the identification of SI.
- Our proposed models have reached efficient accuracy and outperformed the baseline models.

The remainder of the paper is organized as follows. We presented some related studies carried out in the direction of SI in Sect. 2. In Sect. 3, we described our proposed models, the baseline models, in addition to some information about the dataset. Section 4 displays the obtained results and analyzes them in the discussion. Finally, Sect. 5 concludes the paper and clarifies some future directions.

2 Related work

The SI concept is emerging as a common topic in the mental health literature. Various recent studies have proposed different methods to recognize suicide announcements and categorize them into labels [17–20]. Some researchers have implemented Electronic Health Records (EHRs) along with machine learning algorithms to identify the risks of SI [21–23]. Specifically, the patient records involve demographical details and a comprehensive history of diagnoses, including admissions and emergency visits. Furthermore, other researchers have employed machine learning techniques and sentiment analysis for SI classification. For instance, Chiroma et al. [24] have performed Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB) classifiers. They have reached 77.9% accuracy as best binary classification result with DT on Twitter data. Likewise, Sawhney et al. [25] have utilized RF, and Logistic Regression (LR) with feature extraction methods such as Topics Probability, and Term Frequency-Inverse Document Frequency (TF-IDF). The authors have achieved 85.8% accuracy using RF on Twitter data. Shah et al. [26] have implemented NB, SVM, and RF along with hybridized feature extraction methods including N-Gram, TF-IDF, plus Linguistic Inquiry and Word Count (LIWC). The authors proposed framework has obtained 73.6% accuracy through the NB classifier for analyzing Reddit posts. Liu et al. [18] have used a hybrid feature selection approach to monitor

SI from Weibo data and have reached 80.61% accuracy. In practice, they have merged different feature methods such as clustering, word embedding, risk factors for suicide, and basic statistical characteristics. Chatterjee et al. [27] have connected various suggested methods with machine learning classifiers to extract linguistic, statistical, and sentiment analysis features. They have examined their models to catch suicide attempts on Twitter data, and they have obtained 87% accuracy using LR.

Meanwhile, several recent studies have implemented deep-learning algorithms for SI detection. Sawhney et al. [28] have applied Long Short-Term Memory (LSTM), Vanilla Recurrent Neural Network (VRNN), and a C-LSTM architecture that added another convolutional layer to the LSTM model. In effect, the C-LSTM model has promoted the accuracy of detecting suicidal tweets to 81.2%. Sinha et al. [29] have developed an architecture based on text, historical, and Graph Convolutional Networks using BiLSTM model with attention layers. Also, they have experimented CNN and LSTM networks affiliated to the GloVe embedding. Hence, the ensemble modeling methods combined to the BiLSTM classifier has accurately recognized SI tweets by 92.76% using the F1-score measure. Cao et al. [30] have proposed a Suicide-oriented Word Embeddings model based on the Chinese suicide dictionary and the content of Tree Hole dataset. The authors aimed to enrich the sensitivity of suicide-related lexicons and context by capturing the unique linguistic patterns associated with discussions about suicide. Besides, they employed the LSTM model with an attention mechanism alongside their suggested word embedding to identify SI on Weibo dataset. The authors' model demonstrated its effectiveness compared to other standard word embeddings, achieving 91.33% accuracy. Tadesse et al. [31] have performed a set of deep learning algorithms like CNN, LSTM, and a LSTM-CNN that represents a combination of LSTM and CNN architectures with Word2Vec embedding. They have achieved 93.8% accuracy through the LSTM-CNN for the binary classification of suicidal activities on Reddit posts. Ji et al. [32] have concatenated the Vanilla Relation Network plus attention mechanism with LSTM model as a classifier of mental disorder on Reddit and Twitter datasets. The relation network representation has obtained 64.74%, and 83.85% as the best accuracy result on Reddit and Twitter data respectively. After that, Aldhyani et al. [33] have combined CNN and BiLSTM model into a CNN–BiLSTM architecture with Word2Vec embedding to predict whether Reddit posts contained indications of suicidal thoughts or not. The authors model has reached 95% accuracy for the binary classification task. Later, Haque et al. [34] have examined a set of deep learning models including LSTM, BiLSTM, C-LSTM, and Bi-directional Gated Recurrent Unit (BiGRU) to determine suicidal intents on Twitter data. They have attained 93.6% accuracy using BiLSTM model and word embedding technique.

Furthermore, there have been many recent studies that employed transformers and LLMs to detect SI. Haque et al. [35] have implemented various transformer models including BERT, A Lite BERT (ALBERT), Robustly Optimized BERT pretraining approach (RoBERTa), and the generalized autoregressive pretraining method XLNET in order to identify SI in Reddit posts. The BERT-based transformers have shown an efficient accuracy compared to the proposed BiLSTM authors model. In practice, the RoBERTa model has reached 95.21% accuracy as the best authors classification result. Tanaka and Fukazawa [36] employed a framework that combines two LLMs: BERT, which extracts sentences related to suicide risk, and MentaLLaMa (Large Language Model Meta AI), which generates summaries from the identified high-risk sentences. The authors achieved an F1-score of 91.3% for the suicide extraction subtask on Reddit dataset using BERT. Besides, Li et al. [37] have utilized machine learning models alongside the Bio-ClinicalBERT model to predict various suicide tendencies from a manually annotated set of 1000 Initial Psychiatric Evaluation (IPE) notes. The Bio-ClinicalBERT model demonstrated superior performance, achieving a weighted F1-score of 78%. Likewise, Yang et al. [38] have conducted a set of experiments using LLMs for mental health analysis to assess their capabilities and identify areas for improvement. They achieved a 95.11% F1-score for stress classification using the RoBERTa model and obtained an 89.01% F1-score for suicide classification with the MentalRoBERTa model. Ananthakrishnan et al. [39] have utilized several BERT-based LLMs to detect suicidal intentions on Twitter dataset. The authors of RoBERTa model reached the highest accuracy of 96.35%.

In this paper, we propose a C-BiLSTM model using triple word embedding to analyze and detect SI on Reddit posts. As well, we present an analysis of model combinations using CNN, BiLSTM, and pre-trained word embedding GloVe, Word2Vec, and FastText. Additionally, we showcase the potential of employing LLMs, specifically our fine-tuned BERT and GPT models, which have achieved notable classification accuracy. Details on the suggested models are given in the next section.

3 The proposed models

3.1 The dataset

We have used the Reddit dataset from Kaggle, specifically for the purpose of suicide detection via a text classification task [40]. It was sourced from subreddit posts and then categorized into two labels: suicide and no-suicide content. These posts were collected spanning the years from 2008 to

2021, focusing on content related to suicide expressions. The dataset is large and includes 232,074 posts, where the number of posts in both classes is distributed equally. Moreover, 90% of the dataset has an average content length of less or equal to 1653 words.

3.2 Pre-processing and tokenization

Actually, we have prepared our dataset for the training phase by implementing a set of methods to clean Reddit posts. In effect, we have lower cased the text content and cleaned it from numbers and patterns. Besides, we have removed symbols, pictographs, transport and map symbols, as well as emoticons. In addition, we have replaced many contractions in the posts such as “aren’t” with their expanded representations like “are not”. In effect, this process helps to enhance text normalization for more consistent and standardized text data. Also, it facilitates and improves tokenization, which means breaking a text into individual words or tokens [41]. For example, the tokenization of "I’ll" results in a single token ("I’ll"). However, when it is expanded to "I will”, it becomes two separate tokens, "I" and "will". This token expansion can offer advantages in various NLP tasks that require token-level analysis including text classification. After that, we have split our dataset into 75% training data, and 25% validation. Then, we have prepared our input sequence for the training process by tokenizing text data based on word frequency, and padding them all with the same length. The value of the maximum length of the padding parameters and the maximum number of words in the tokenization process are presented in “The Architecture” subsection.

3.3 Word embedding

Primarily, word embedding is a necessary element of NLP that converts words into high-dimensional vectors of representation values. The vectors acquire contextual information about words and semantic relationships between them, allowing better language processing and understanding. In the main, three word embedding pre-trained models are commonly used in text classification tasks:

- **Word2Vec** [42]: is an embedding representation that involves 300 similarity vector values for 1.6 billion words. Word2Vec is able to capture the analogy relationships and words’ similarity to encode the semantic meanings efficiently.
- **GloVe** [43]: represents a 300 dimensional vector of word co-occurrence statistics for 840 billion tokens. It performs well at extracting word semantics for word analogy tasks.
- **FastText** [44]: is a pre-trained embedding model that contains 300 dimension word vectors of 2 million words. It enriches the embeddings by considering subword information, which allows to effectively manage the morphological nuances between words.

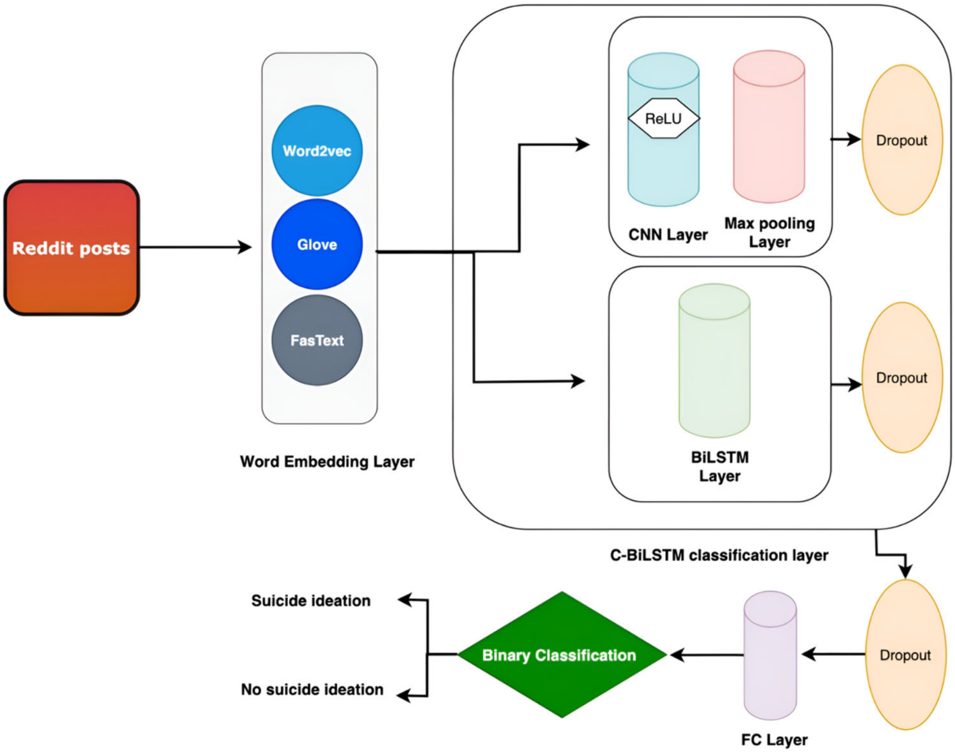
In practice, we have examined this three word embedding models with different deep learning models such as CNN, BiLSTM, and C-BiLSTM. First, we have experimented the three word embeddings using CNN and BiLSTM. Next, we linked the CNN model to a double word embedding via Word2Vec and GloVe to increase the performance of the posts classification. Then, we have assessed a merged deep learning version C-BiLSTM with a dual representation of the three embeddings by generating all the conceivable combinations from them. Finally, we have built a developed architecture including the C-BiLSTM along with a triple word embedding representation through Word2Vec, Glove, and FastText. The description and the results of these proposed models are displayed subsequently.

3.4 The architecture

The model architecture presented in Fig. 1 is a CNN with an integrated BiLSTM layer used for the text classification task. It is designed to handle SI from Reddit text data by employing three different pre-trained word embeddings: Word2Vec, FastText, and GloVe. The model represents our best proposed deep learning method among various experimented CNN, and BiLSTM models with different word embedding combinations. First of all, we have started to load each of the three word embedding sources and combined them along the feature dimension. Specifically, we have set the embedding size to 300 in order to match all the pre-trained word embedding dimensions. Besides, we have specified the maximum feature to 120.000 unique words, and 800 maximum length of each input sentence. Then, we have passed the embeddings through the convolutional layer using 1, 2, 3, and 5 as filter size values. This convolutional layer employed 100 filters allowing the model to catch several n-gram patterns in the text. Also, we have applied the Rectified Linear Unit (ReLU) activation function to introduce non-linearity on data, and a dropout to prevent overfitting. Next, the output has been provided to the BiLSTM layer with 128 hidden units. In effect, the BiLSTM layer captured contextual feature in both forward and backward directions. As well, we have applied a dropout with a rate of 0.7 to prevent overfitting. Furthermore, the C-BiLSTM classifier layer with a dropout have been linked with a fully connected linear layer with the number of output classes. Finally, the model outputted the logits for each class, which are then used to compute the cross-entropy loss during training.

Below is a simplified representation of the operations in each layer of our model:

Fig. 1 The proposed architecture for the C-BiLSTM model



a) Embedding Layers

- GloVe Embedding

$$x\_glove[i, j] = E\_glove[word\_idx] \quad (1)$$

where **"i"** is the batch index, **"j"** is the position in the sequence, **"E_glove"** is the GloVe embedding matrix, and **"word_idx"** is the index of the word in the vocabulary.

- Word2Vec Embedding

$$x\_word2vec[i, j] = E\_word2vec[word\_idx] \quad (2)$$

Using the same embedding notation as GloVe Embedding.

- FastText Embedding

$$x\_fasttext[i, j] = E\_fasttext[word\_idx] \quad (3)$$

Using the same embedding notation.

b) Concatenation of Embeddings

$$x\_combined[i, j] = [x\_glove[i, j], x\_word2vec[i, j], x\_fasttext[i, j]] \quad (4)$$

The concatenation of the three embeddings along the features dimension.

C) CNN Layers

- Convolutional layer
For each filter size **"K",** we applied 2D convolution with a filter **"W_K"** and a ReLU activation function:

$$conv\_result\_K[i, j] = ReLU(conv(x\_combined[i, j : j + K], W\_K) + b\_K) \quad (5)$$

where **"conv"** denotes the convolution operation, "**x_combined[i,j:j + K]"** represents the input sequence of length **"K"** where **"j"** is the starting position, and **"b_K"** is the bias term associated with the filter **"W_K"**.

- Max-pooling layer

$$max\_p\_K[i] = Max(conv\_result\_K[i, j]) \quad (6)$$

where **"Max"** is the maximum value over the temporal dimension.

- Combined filter sizes

$$x1[i] = Concat(max\_p\_1[i], max\_p\_2[i], max\_p\_3[i], max\_p\_5[i]) \quad (7)$$

where 1, 2, 3, and 5 represent the filter sizes.

d) BiLSTM Layer: we use a Bidirectional LSTM

- Forward LSTM

$$LSTM\_fwd[i,\, j] = LSTM\_fw(x\_combined[i,\, j]) \tag{8}$$

“LSTM_fw” is the LSTM forward pass.

- Backward LSTM

$$LSTM\_bwd[i,\, j] = LSTM\_bw(x\_combined[i,\, j]) \tag{9}$$

“LSTM_bw” is the LSTM backward pass.

- Combined hidden states

$$x\_bilstm[i] = Concat(LSTM\_fwd[i,\, -1],\, LSTM\_bwd[i,\, 0]) \tag{10}$$

where **“-1”** and **“0”** represent the last and first positions in the sequence.

e) Dropout layers

$$x1\,[i] = Dropout\,(x1\,[i])\, and x\_bilstm[i] = Dropout\,(x\_bilstm\,[i]) \tag{11}$$

where **“ x 1[i]”** is the combined output of CNN layers and **“x_bilstm[i]”** is the combined output of BiLSTM layers.

f) C-BiLSTM layer with dropout

$$x[i] = Dropout(Concat(x1[i],\, x\_bilstm[i])) \tag{12}$$

where **“x[i]”** is the combined tensor of **“ x 1[i]”** and **“x_bilstm[i]”**.

g) Fully connected layer

$$logit[i] = FC(x[i]) \tag{13}$$

“FC” is a fully connected layer with weights and biases applied to **“x[i]”** to produce the logit scores for each class.

3.5 Other architectures and the baseline models

Basically, we have trained several architecture models to assess their effect on the classification task, as well as developing the accuracy of our proposed models below:

- **CNN model**: involves two channels of convolutional layer. Each one employed 36 filters through 1, 2, 3, and 5 filter sizes, and affiliated with a dropout layer and a max-pooling layer. The max-pooled outputs from both sets of convolutional layers have been merged together to form a single feature vector that has been passed to a fully connected layer. In effect, we have evaluated the three pre-defined word embedding models with CNN to analyze their impact on the performance of the model.
- **BiLSTM model**: is a category of Recurrent Neural Network (RNN) that manages the input text in both forward and backward directions. It contains a BiLSTM layer with a hidden size of 64 units connected to a linear layer with the ReLU activation function, and connected to a dropout layer for regularization. Subsequently, we have applied average and max pooling operations to the LSTM outputs and concatenated them before passing them through the linear layer. Then, the output layer produced the final predictions for each class as suicidal thoughts or not. Moreover, we have compared the results of each affiliated word embedding model to the BiLSTM in the classification task, following the same approach as the CNN model.
- **C-BiLSTM model**: is the model pre-described in the previous subsection. Actually, the model has been performed using a double combination of the three word embedding models including all possible compositions. Afterwards, we have implemented a triple word embedding technique that merged them to get one feature vector.

After performing deep learning models, we have utilized various types of BERT LLMs, including Bio-ClinicalBERT [45], MentalBERT [46], MentalRoBERTa [46], in addition to the GPT2 transformer model [47]. Each BERT model has been pre-trained on extensive health datasets. For instance, Bio-ClinicalBERT has been trained on the PubMed dataset, while MentalBERT and MentalRoBERTa have been pre-trained using the mental health of subreddit posts. GPT2, on the other hand, is a transformer model pre-trained on a vast corpus of English data. In practice, we fine-tuned each model using various hyperparameter values to achieve optimal performance in identifying SI from social media posts. The results of each model are presented and discussed in the following section.

Furthermore, we have compared our top-performing models with a set of deep learning, machine learning, and transformer models. Indeed, all models have been implemented to identify SI in Reddit social media posts.

- **NB** [26]: this model combines different feature extraction techniques including n-gram, and TF-IDF, in addition to the linguistic feature analysis LIWC. This hybridized feature approach has been implemented along with the NB classifier providing the best author’s architecture for the classification of SI.
- **SVM** [48]: represents a set of extracted features from Reddit posts using algorithms such as TF-IDF, and LIWC,

along with sentiment analysis. Then, these linguistic and semantic features have been incorporated into the SVM classifier to identify and detect suicidal thoughts.

- **Relation Network (RN)** [32]: is a new LSTM model which includes the vanilla relation network and attention. In practice, the relation network module computes scores that represent the relationships between individual tokens in posts and state indicators. These state indicators are identified by sentiment features. Then, the model assigned the attention weights to the acquired relationships within the relation module. As well, the LSTM network extracted sequential independencies from the encoded text representation and the state indicators. Lastly, the classifier took the learned representation and provided a class prediction.
- **ACL (Attention over Convolution and Long short-term memory neural network)** [49]: is a merged version of CNN, LSTM, and attention mechanism. Actually, the model employed the GloVe embedding to the input text followed by a convolutional layer and a max pooling layer. The output has been fed to a BiLSTM layer and then forwarded to the attention layer for more specific features learning. Finally, two dense layers obtained the output result and predicted the final class output through the sigmoid activation function.
- **CNN-LSTM** [31]: the proposed model implemented a Word2Vec embedding layer affiliated with a dropout layer to avoid overfitting. After that, a LSTM layer acquired contextual informations from the embedding layer, and learned long-distance dependencies. Also, a convolutional layer with a max pooling layer learned more contextual features and provided the output to the SoftMax layer to produce the final classification result of the model.
- **BERT:** is a fine-tuned model utilized by Tanaka and Fukazawa [36], set up with hyperparameters with a batch size of 8, a learning rate of 2e-5, and a warm-up ratio of 0.1. The transformer model carried out training for 50 epochs to classify suicidal thoughts in Reddit dataset as negative or positive.
- **RoBERTa**: is a variant of the BERT model, pre-trained with longer sequences and on a larger dataset. It was implemented by Haque et al. [35] for SI detection on Reddit dataset.

3.6 The training parameters and metrics

Concerning the training parameters, our deep learning models have been trained for 10 epochs using a batch size of 64 samples. Moreover, we have optimized the parameters using Adam optimizer with a learning rate of 0.001. To handle the learning rate, a step-wise learning rate scheduler has been employed, so as to reduce its value by a factor of 0.6 every 2 epochs. Meanwhile, for LLMs, we trained our models loaded from the Hugging Face library [50] for 3 epochs with learning rates of 1e-3, 1e-5, 1e-6, and 1e-7. Additionally, we experimented with different batch sizes, including 16, 32, and 64. In practice, selecting the optimal parameters is the result of multiple training iterations and outcomes monitoring. First, we initiated our training with GPT2 using a batch size of 64 and a learning rate of 1e-5. Subsequently, we examined different values to improve the SI detection results, including adjusting the dropout rate to prevent overfitting in certain scenarios. The hyperparameter values along with their corresponding results are presented in the following section. In effect, we trained all our models on Kaggle, which provides free access to NVIDIA TESLA P100 GPUs and 16GB of RAM. During training, the models have been evaluated on both the training and validation datasets. For each epoch, we have computed the training and validation losses, as well as the accuracy (14) and the weighted F1-score (17) metrics. Also, we have monitored the training progress to identify prospective signs of overfitting or underfitting. Explicitly, "TP" represents "True Positive" which is the model's correct prediction of the positive class. Also, "TN" is "True Negative" that represents the model's correct prediction of the negative class. "FP" is "False Positive" which refers to the model's incorrect prediction of the positive class. Besides, "FN" is "False Negative" that signifies the model's incorrect prediction of the negative class [51].

- **Accuracy** [51]: provides a straightforward measure of a model's performance by quantifying the ratio of correctly predicted instances to the entire instances in the dataset. It offers a clear and intuitive evaluation of its overall efficiency in making predictions particularly for symmetric datasets.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{14}$$

- **Precision** [51]: focuses on the count of true positive predictions against the entire positive predictions produced by the model. It serves as a crucial metric in evaluating a model's capacity to make accurate positive predictions while minimizing false positives. A high precision value means that the model exhibits a low incidence of false positive predictions, underlining its precision and reliability in correctly identifying positive cases.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

- **Recall (Sensitivity)** [51]: considers the count of true positive predictions against the entire actual positives in the dataset. Which means it evaluates the model's accuracy in identifying all the positive instances in a dataset. When the

Table 1 The SI detection results of our deep learning models

Model	Accuracy	F1-score
FastText-BiLSTM	90.04	90.04
Word2Vec-BiLSTM	90.14	90.13
GloVe-BiLSTM	90.83	90.83
FastText-CNN	91.08	91.08
Word2Vec-CNN	91.41	91.41
GloVe-CNN	91.76	91.76
Double-embed-C-BiLSTM (w–f)	91.19	91.18
Double-embed-C-BiLSTM (f-g)	91.77	91.77
Double-embed-C-BiLSTM (w-g)	92.86	92.86
Triple-embed-C-BiLSTM	**94.95**	**94.95**

The model with the best performance is set to bold

recall is high, it signifies that the model has a strong capability to minimize false negatives, effectively capturing a substantial portion of positive cases, and demonstrating its performance in reducing the number of missed positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

- **F1-score** [51]: represents the weighted average of Recall and Precision. In its calculation, FP and FN are both considered, making it a valuable metric for asymmetric datasets. Unlike accuracy, which may not adequately capture performance imbalances, the F1-score offers a more insightful evaluation of a model's ability to balance Precision and Recall, making it a preferred choice for assessing the classification performance of asymmetric datasets.

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{17}$$

4 Results and discussion

Actually, we have used advanced deep learning models and LLMs to analyze and extract SI from Reddit social media posts. The best empirical deep learning results of our text classification analysis are displayed in Table 1 considering the accuracy and F1-score metrics. Explicitly, we refer to Word2Vec in the double embedding representation by "w", GloVe by "g", and FastText by "f". Notably, the C-BiLSTM model including triple word embedding has attained the best accuracy and F1-score result among deep learning models by 94.95%. Figures 2 and 3 represent the loss and accuracy curves, as well as the confusion matrix achieved by the C-BiLSTM model with triple embeddings, respectively. These results highlight the potential advantages of using a triple features concatenation in pre-trained embedding models. In effect, reached results clearly display that our models' performance improved as we integrated additional sources of word embeddings. Practically, each added source, whether it was FastText, Glove, or Word2Vec, contributes in enhancing the model accuracy and F1-score. This observation highlights the synergistic effect of combining diverse word embeddings, suggesting that leveraging multiple embedding sources can lead to superior performance in text classification tasks.

On the other hand, we can notice that the CNN models' performance exceeds those of BiLSTM models. Essentially, CNN has consistently outperformed BiLSTM across all word embedding options, achieving more efficient classification results. This high performance can be due to the convolutional filters' ability to capture meaningful features from the embedding models, especially in sentiment analysis problems. In contrast, the BiLSTM model has also shown its performance as a powerful deep-learning algorithm for sentiment analysis tasks. In practice, BiLSTM can consider both past and future contexts for each word in a sentence, providing a deeper understanding of text sentiments. It significantly enhanced the accuracy of the CNN model in the C-BiLSTM version, leveraging their complementary abilities in feature extraction and sequence modeling, respectively. Besides, we can notice that the C-BiLSTM model has showed their capacity of procuring more contexts from the largest representation of word embedding. It has reached 92.86% accuracy as the best double incorporation model using a Word2Vec and GloVe embedding combination.

Alternatively, the comparison of the three pre-trained word embedding results reveals that GloVe embeddings have performed slightly better than Word2Vec and FastText embeddings. In effect, the larger dimensionality of GloVe embeddings compared to Word2Vec or FastText, allows to capture more nuanced and frequent word semantics within the text data. Hence, GloVe embeddings could be an optimal choice within sentiment analysis or similar tasks, offering distinct advantages in enhancing a classification model performance.

For the LLMs, we conducted various experiments and monitored the results using different hyperparameter values. Table 2 displays our attempts to adjust the parameters of each model and the corresponding results achieved. In practice, we fine-tuned the models by modifying the batch size, learning rate (LR), and dropout rate values. We could notice that each model has different optimal hyperparameter values. For instance, GPT2 achieved its best performance with a batch size of 16, an LR of 1e-6, and a dropout rate of 0.1. Meanwhile, Bio-ClinicalBERT demonstrated good performance without the need for dropout, and with batch size and LR values similar to those of GPT2. MentalBERT and

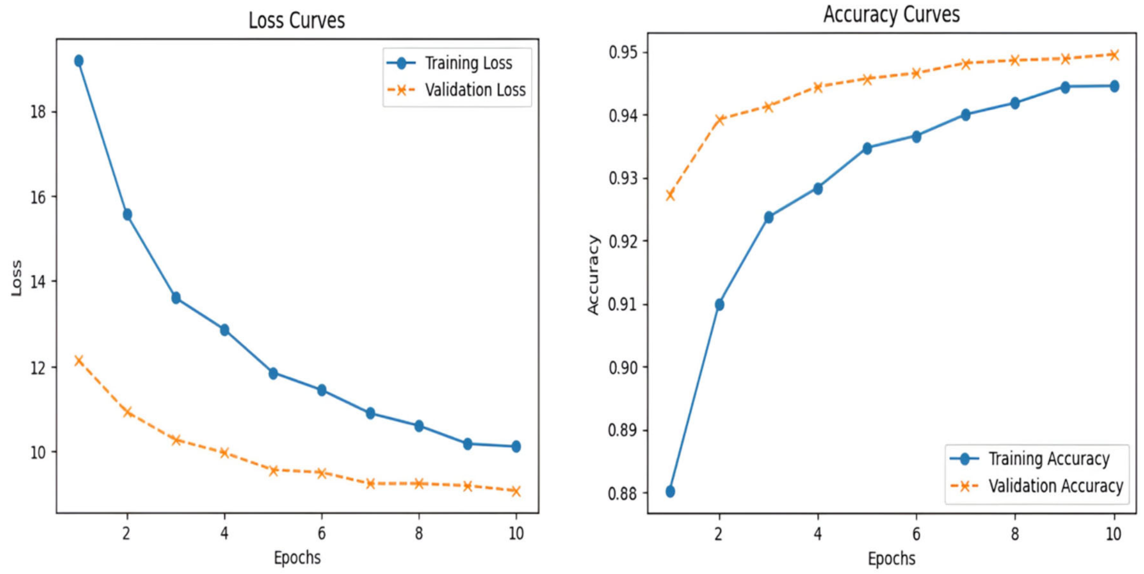


Fig. 2 The loss and the accuracy curves of our best C-BiLSTM model with triple embeddings

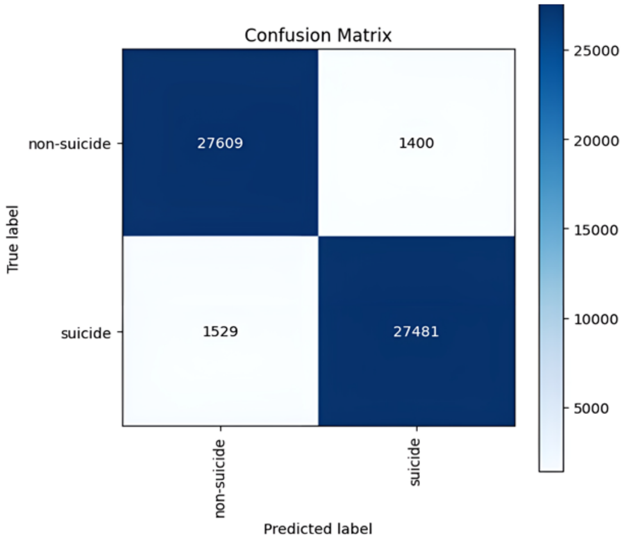


Fig. 3 The confusion matrix of our best C-BiLSTM model with triple embeddings

MentalRoBERTa achieved appropriate results with an LR of 1e-7, a batch size of 16, and a dropout rate of 0.1.

Nevertheless, our optimal GPT2 model surpassed all other LLMs and deep learning models, achieving an accuracy and F1-score of 97.69%. Figure 4 illustrates the accuracy curves and the result obtained by the GPT2 model. This notable superiority over deep learning methods underscores the effectiveness of utilizing LLMs in sentiment analysis and NLP tasks. Indeed, the transformer architecture, exemplified in BERT-based and GPT2 models, facilitates the comprehensive understanding of contextual information and intricate relationships between words within sentences in datasets. Moreover, the self-attention mechanism allows LLMs to efficiently process long-range dependencies and capture global contextual information, further enhancing their performance in complex NLP tasks. Additionally, LLMs exhibit remarkable adaptability and generalization capabilities due to their ability to dynamically adjust to different domains and languages, including the mental health domain. As a result, they demonstrate robust performance by leveraging the extensive pre-training Reddit social media posts dataset that aids in capturing syntactic structures, linguistic patterns, and semantic meanings.

Furthermore, Table 3 demonstrates that our top-performing LLM, and our best deep learning model outperform the baseline models. Indeed, using a triple merged representation of word embedding with the C-BiLSTM architecture has promoted the accuracy of SI detection surpassing all the pre-presented machine learning and deep learning models. For instance, the RN [32] utilized sentiment lexicons and topic modeling to enhance text representation. However, this inadvertently impacted the model performance due to propagation of errors during the data preprocessing stage. In fact, inaccuracies in sentiment analysis or topic modeling may be due to the biased data representation. Besides, the RN's reliance on lexicon-based sentiment scores and latent topics may not adequately capture the complex nuances of language and sentiment, particularly in a complex domain like mental health. In contrast, ACL [49] utilized Word2Vec techniques for dataset preprocessing and employed various types of word embeddings. The authors also tuned the hyperparameters using a grid search to optimize the performance of the ACL model. Hence, ACL surpassed the limitations observed in the RN [32] model,

Table 2 The proposed LLMs' hyper-parameters and reached results on SI detection

Model	LR	Batch size	Dropout	Accuracy	F1-score	Overfitting
GPT2	1e-5	64	–	78.09	78.11	no
	1e-3	32	–	91	91	no
	1e-7	16	–	89.26	88.06	yes
	1e-6	**16**	**0.1**	**97.69**	**97.69**	**no**
Bio-ClinicalBERT	**1e-6**	**16**	–	**96.91**	**96.94**	**no**
MentalBERT	1e-6	16	–	97.72	97.72	yes
	1e-6	16	0.1	97.70	97.70	yes
	1e-6	16	0.3	97.38	97.43	yes
	1e-6	16	0.5	97.66	97.67	yes
	1e-7	**16**	**0.1**	**95.66**	**95.71**	**no**
MentalRoBERTa	1e-6	16	0.1	97.70	97.71	yes
	1e-6	16	0.7	98.61	98.60	yes
	1e-6	16	0.6	98.80	98.81	yes
	1e-7	**16**	**0.1**	**93.30**	**93.42**	**no**

The optimal performance values of each model are set to bold

achieving an F1-score of 90.82%. Additionally, the utilization of hybrid classification models, such as merging CNN and BiLSTM in ACL [49] and CNN-LSTM [31] models, significantly enhanced their ability to comprehend complex linguistic patterns associated with SI. Consequently, our deep learning Triple-embed-C-BiLSTM model effectively mitigates the baseline models' shortcomings and improves text classification performance. In addition, our fine-tuned LLMs have significantly improved the performance of mental health analysis, yielding notable results. Specifically, our GPT2 model achieved an accuracy and F1-score of 97.69%, surpassing all the baseline performance. In fact, the GPT2 leverages the robust architecture of transformers, along with the inherent features pre-learned from extensive datasets. These advantages render large models powerful tools for SI detection, leading to more accurate predictions and analysis.

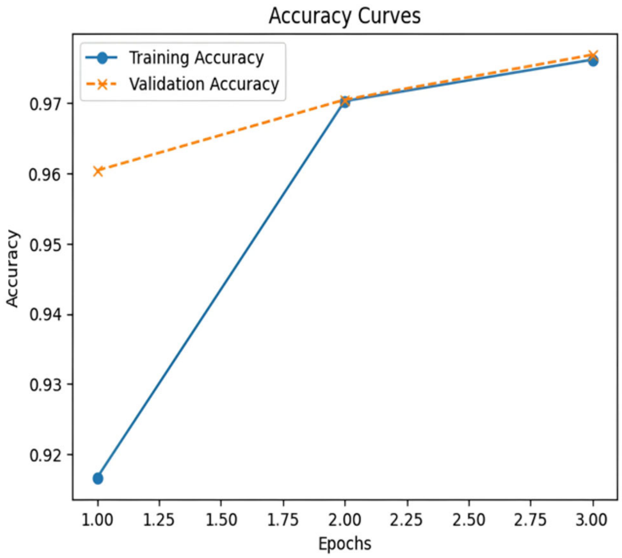


Fig. 4 The accuracy curves of our GPT2 model

5 Conclusion and future work

This paper introduces various deep learning architectures and several LLMs to detect suicidal thoughts in Reddit posts. In practice, we employed CNN and BiLSTM models along with multiple pre-trained word embedding models. Through the combination of CNN and BiLSTM models, we enhanced mental health analysis. In fact, our hybrid C-BiLSTM classification model, leveraging three pre-trained word embeddings, achieved 94.95% accuracy and F1-score, while addressing several limitations in other deep learning and Large Language baseline models. By extensively incorporating words learned from merged word embeddings, we effectively surpass other methods like topic modeling and sentiment lexicons. Furthermore, the merging of CNN and BiLSTM models enables the extraction of richer features, harnessing the potential of filters and the attention mechanism. Additionally, our LLMs outperform traditional deep learning methods, achieving remarkable sentiment analysis performance. Specifically, our GPT2 model attained 97.69% accuracy and F1-score, benefiting from GPT's extensive data and transformer architecture. Looking ahead, our focus lies on refining our proposed models for optimal performance across diverse datasets, ultimately aiming to prevent suicide and offer support to individuals in distress.

Table 3 Performance comparison of the proposed models with baselines

Model		Accuracy	F1-score
Machine learning	NB [26]	73.6	76.7
	SVM [48]	92	92
Deep learning	RN [32]	64.74	64.78
	ACL [49]	88.48	90.82
	CNN-LSTM [31]	93.8	93.4
	Triple-embed-C-BiLSTM	**94.95**	**94.95**
LLMs	BERT [36]	–	91.3
	RoBERTa [35]	95.21	–
	GPT2	**97.69**	**97.69**

The model with the best performance is set to bold

References

1. Naguy, A., Elbadry, H., Salem, H.: Suicide: a précis! J. Fam. Med. Prim. Care **9**, 4009–4015 (2020). https://doi.org/10.4103/jfmpc.jfmpc_12_20
2. Klonsky, E.D., May, A.M., Saffer, B.Y.: Suicide, suicide attempts, and suicidal ideation. Annu. Rev. Clin. Psychol. **12**, 307–330 (2016). https://doi.org/10.1146/annurev-clinpsy-021815-093204
3. Jobes, D.A., Joiner, T.E.: Reflections on suicidal ideation. Crisis **40**, 227–230 (2019). https://doi.org/10.1027/0227-5910/a000615
4. Norris, D.R., Clark, M.S.: The suicidal patient: evaluation and management. Am. Fam. Phys. **103**, 417–421 (2021). (**PMID: 33788523**)
5. Rodriguez, C.I., Zorumski, C.F.: Rapid and novel treatments in psychiatry: the future is now. Neuropsychopharmacology **48**, 1–2 (2023). https://doi.org/10.1038/s41386-023-01720-2
6. Kennard, B.D., Hughes, J.L., Minhajuddin, A., et al.: Suicidal thoughts and behaviors in youth seeking mental health treatment in Texas: youth depression and suicide network research registry. Suicide Life-Threatening Behav. **53**, 748–763 (2023). https://doi.org/10.1111/sltb.12980
7. Hophing, M., Zimmerman-Winslow, K.J., Basu, A., Jacob, T.: The impact of COVID-19 and quarantine on suicidality in geriatric inpatients—a case Report. J. Geriatr. Psychiatr. Neurol. **35**, 550–554 (2022). https://doi.org/10.1177/08919887211023588
8. Hinduja, S., Patchin, J.W.: Bullying, cyberbullying, and suicide. Arch. Suicide Res. **14**, 206–221 (2010). https://doi.org/10.1080/13811118.2010.494133
9. Hong, J.S., Kral, M.J., Sterzing, P.R.: Pathways from bullying perpetration, victimization, and bully victimization to suicidality among school-aged youth: a review of the potential mediators and a call for further investigation. Trauma Violence Abus. **16**, 379–390 (2015). https://doi.org/10.1177/1524838014537904
10. Vergara, G.A., Stewart, J.G., Cosby, E.A., et al.: Non-Suicidal self-injury and suicide in depressed adolescents: impact of peer victimization and bullying. J. Affect. Disord. **245**, 744–749 (2019). https://doi.org/10.1016/j.jad.2018.11.084
11. Nobles, A.L., Glenn, J.J., Kowsari, K., et al.: Identification of Imminent suicide risk among young adults using text messages. In: CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp 1–11. https://doi.org/10.1145/3173574.3173987(2018)
12. Gaur, M., Kursuncu, U., Sheth, A., et al.: Knowledge-aware assessment of severity of suicide risk for early intervention. In: The Web Conference 2019-Proceedings of the World Wide Web Conference, WWW 2019. Association for Computing Machinery, New York, NY, USA, pp 514–525. https://doi.org/10.1145/3308558.3313698(2019)
13. Tseng, F.Y., Yang, H.J.: Internet use and web communication networks, sources of social support, and forms of suicidal and non-suicidal self-injury among adolescents: different patterns between genders. Suicide Life-Threatening Behav. **45**, 178–191 (2015). https://doi.org/10.1111/sltb.12124
14. Lopez-Castroman, J., Moulahi, B., Azé, J., et al.: Mining social networks to improve suicide prevention: a scoping review. J. Neurosci. Res. **98**, 616–625 (2020). https://doi.org/10.1002/jnr.24404
15. Wang, Z., Yu, G., Tian, X.: Exploring behavior of people with suicidal ideation in a Chinese online suicidal community. Int. J. Environ. Res. Public Health **16**, 54–67 (2019). https://doi.org/10.3390/ijerph16010054
16. Yue, L., Chen, W., Li, X., et al.: A survey of sentiment analysis in social media. Knowl. Inf. Syst. **60**, 617–663 (2019). https://doi.org/10.1007/s10115-018-1236-4
17. Diniz, E.J.S., Fontenele, J.E., de Oliveira, A.C., et al.: Boamente: a natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. Healthcare **10**, 698–717 (2022). https://doi.org/10.3390/healthcare10040698
18. Liu, J., Shi, M., Jiang, H.: Detecting suicidal ideation in social media: an ensemble method based on feature fusion. Int. J. Environ. Res. Public Health **19**, 8197–8210 (2022). https://doi.org/10.3390/ijerph19138197
19. Chadha, A., Gupta, A., Kumar, Y.: Suicidal ideation detection on social media: a machine learning approach. In: Proceedings of International Conference on Technological Advancements in Computational Sciences, ICTACS 2022. IEEE, Tashkent, Uzbekistan, pp 685–688. (2022) https://doi.org/10.1109/ICTACS56270.2022.9988722
20. Chadha, A., Kaushik, B.: A survey on prediction of suicidal ideation using machine and ensemble learning. Comput. J. **64**, 1617–1632 (2021). https://doi.org/10.1093/comjnl/bxz120
21. Van, L.D., Montgomery, J., Kirkby, K.C., Scanlan, J.: Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. J. Biomed. Inform. **86**, 49–58 (2018). https://doi.org/10.1016/j.jbi.2018.08.007
22. Bittar, A., Velupillai, S., Roberts, A., Dutta, R.: Text classification to inform suicide risk assessment in electronic health records. Stud. Health Technol. Inform. **264**, 40–44 (2019). https://doi.org/10.3233/SHTI190179
23. Carson, N.J., Mullin, B., Sanchez, M.J., et al.: Identification of suicidal behavior among psychiatrically hospitalized adolescents

using natural language processing and machine learning of electronic health records. PLoS ONE **14**, e0211116 (2019). https://doi.org/10.1371/journal.pone.0211116

24. Chiroma, F., Liu, H., Cocea, M.: Text classification for suicide related tweets. In: Proceedings-International Conference on Machine Learning and Cybernetics. IEEE, Chengdu, China, pp 587–592. (2018) https://doi.org/10.1109/ICMLC.2018.8527039
25. Sawhney, R., Manchanda, P., Singh, R., Aggarwal, S.: A computational approach to feature extraction for identification of suicidal ideation in tweets. In: ACL 2018-56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop. Association for Computational Linguistics, Melbourne, Australia, pp 91–98. (2018) https://doi.org/10.18653/v1/p18-3013
26. Shah, F.M., Haque, F., Un Nur, R., et al.: A hybridized feature extraction approach to suicidal ideation detection from social media post. In: 2020 IEEE Region 10 Symposium, TENSYMP 2020. IEEE, Dhaka, Bangladesh, pp 985–988. (2020) https://doi.org/10.1109/TENSYMP50017.2020.9230733
27. Chatterjee, M., Samanta, P., Kumar, P., Sarkar, D.: Suicide Ideation detection using multiple feature analysis from twitter data. In: 2022 IEEE Delhi Section Conference, DELCON 2022. IEEE, New Delhi, India, pp 1–6. (2022) https://doi.org/10.1109/DELCON54057.2022.9753295
28. Sawhney R, Manchanda P, Mathur P, et al.: Exploring and Learning suicidal ideation connotations on social media with deep learning. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, Brussels, Belgium, pp 167–175. (2019) https://doi.org/10.18653/v1/w18-6223
29. Sinha, P.P., Mahata, D., Mishra, R., et al.: #suicidal-A multipronged approach to identify and explore suicidal ideation in twitter. In: International Conference on Information and Knowledge Management, Proceedings. Association for Computing Machinery New York, NY, United States, pp 941–950. (2019) https://doi.org/10.1145/3357384.3358060
30. Cao, L., Zhang, H., Feng, L., et al.: Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 1718–1728. (2019) https://doi.org/10.18653/v1/D19-1181
31. Tadesse, M.M., Lin, H., Xu, B., Yang, L.: Detection of suicide ideation in social media forums using deep learning. Algorithms **13**, 7–26 (2020). https://doi.org/10.3390/a13010007
32. Ji, S., Li, X., Huang, Z., Cambria, E.: Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput. Appl. **34**, 10309–10319 (2022). https://doi.org/10.1007/s00521-021-06208-y
33. Aldhyani, T.H.H., Alsubari, S.N., Alshebami, A.S., et al.: Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. Int. J. Environ. Res. Public Health **19**, 12635–12651 (2022). https://doi.org/10.3390/ijerph191912635
34. Haque, R., Islam, N., Islam, M., Ahsan, M.M.: A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning. Technologies **10**, 57–72 (2022). https://doi.org/10.3390/technologies10030057
35. Haque, F., Nur, R.U., Jahan, S. Al, et al.: A transformer based approach to detect suicidal ideation using pre-trained language models. In: ICCIT 2020—23rd International Conference on Computer and Information Technology, Proceedings. IEEE, DHAKA, Bangladesh, pp 1–5. (2020) https://doi.org/10.1109/ICCIT51783.2020.9392692
36. Tanaka, R., Fukazawa, Y.: Integrating supervised extractive and generative language models for suicide risk evidence summarization. In: CLPsych 2024 - 9th Workshop on Computational Linguistics and Clinical Psychology, Proceedings of the Workshop. Association for Computational Linguistics, St. Julians, Malta, pp 270–277. (2024) https://doi.org/10.48550/arXiv.2403.15478
37. Li, Z., Ameer, I., Hu, Y., et al.: Suicide tendency prediction from psychiatric notes using transformer models. In: Proceedings - 2023 IEEE 11th International Conference on Healthcare Informatics, ICHI 2023. IEEE Computer Society, Los Alamitos, CA, USA, pp 481–483. (2023) https://doi.org/10.1109/ICHI57859.2023.00074
38. Yang, K., Ji, S., Zhang, T., et al.: Towards interpretable mental health analysis with large language models. In: EMNLP 2023—2023 Conference on Empirical Methods in Natural Language Processing, Proceedings. The Association for Computational Linguistics, Singapore, pp 6056–6077. (2023) https://doi.org/10.18653/v1/2023.emnlp-main.370
39. Ananthakrishnan, G., Jayaraman, A.K., Trueman, T.E., et al.: Suicidal intention detection in tweets using BERT-based transformers. In: 3rd IEEE 2022 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2022. IEEE, Greater Noida, India, pp 322–327. (2022) https://doi.org/10.1109/ICCCIS56430.2022.10037677
40. Suicide Detection Dataset. https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch. Accessed 28 Sep 2023
41. Hickman, L., Thapa, S., Tay, L., et al.: Text preprocessing for text mining in organizational research: review and recommendations. Organ. Res. Methods **25**, 114–146 (2022). https://doi.org/10.1177/1094428120971683
42. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings. ICLR 2013 - Workshop Track Proceedings. https://doi.org/10.48550/arXiv.1301.3781. (2013) Accessed 28 Sep 2023
43. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: EMNLP 2014—2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. Association for Computational Linguistics, Doha, Qatar, pp 1532–1543. (2014) https://doi.org/10.3115/v1/d14-1162
44. Mikolov, T., Grave, E., Bojanowski, P., et al.: Advances in pre-training distributed word representations. In: LREC 2018—11th International Conference on Language Resources and Evaluation. pp 52–55. (2019) https://doi.org/10.48550/arXiv.1712.09405. Accessed 28 Sep 2023
45. Alsentzer, E., Murphy, J., Boag, W. et al.: Publicly Available Clinical {BERT} Embeddings. In: Rumshisky A, Roberts K, Bethard S, Naumann T (eds) Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp 72–78. (2019) https://doi.org/10.18653/v1/W19-1909
46. Ji, S., Zhang, T., Ansari, L. et al.: MentalBERT: Publicly available pretrained language models for mental healthcare. In: 2022 Language Resources and Evaluation Conference, LREC 2022. European Language Resources Association, Marseille, France, pp 7184–7190. (2022) https://doi.org/10.48550/arXiv.2110.15621
47. Radford, A., Wu, J., Child, R., et al.: Language models are unsupervised multitask learners. https://api.semanticscholar.org/CorpusID:160025533. (2019) Accessed 29 Apr 2024
48. Aladag, A.E., Muderrisoglu, S., Akbas, N.B., et al.: Detecting suicidal ideation on forums: proof-of-concept study. J. Med. Internet Res. **20**, e215 (2018). https://doi.org/10.2196/jmir.9840
49. Chadha, A., Kaushik, B.: A hybrid deep learning model using grid search and cross-validation for effective classification and prediction of suicidal ideation from social network data. New Gener.

Comput. **40**, 889–914 (2022). https://doi.org/10.1007/s00354-022-00191-1

50. Hugging Face. https://huggingface.co/. Accessed 29 Apr 2024
51. Nabeel, M., Rehman, A., Shoaib, U.: Accuracy Based feature ranking metric for multi-label text classification. Int. J. Adv. Comput. Sci. Appl. **8**, 369–379 (2017). https://doi.org/10.14569/ijacsa.2017.081048

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.