Title: A Lexicon-based Approach to Detecting Suicide-related messages on Twitter

Article Type: Research Paper

Keywords: Twitter; suicidal thoughts; mental health; lexicon-based approach; semi-supervised learning; incidents detection

Corresponding Author: Dr. HOSAM ALSAMARRAIE,

First Author: Samer M Sarsam

Order of Authors: Samer M Sarsam; Hosam Al-Samarraie; Ahmed Ibrahim Alzahrani; Waleed S Alnumay; Andrew P Smith

Abstract: Expression of emotion is an indicator that can contribute to the detection of mental health-related disorders. Suicide causes death to many people around the globe, and despite the suicide prevention strategies that have been employed over the years, only a few studies have explored the role of emotions in predicting suicidal behavior on social media platforms. This study explored the role of emotions from Twitter messages in detecting suicide-related content. We extracted and analyzed the characteristics of Twitter users' sentiment and behavior response (anger, fear, sadness, joy, positive, and negative) using NRC Affect Intensity Lexicon and SentiStrength techniques. A semi-supervised learning method was applied using the YATSI classifier or "Yet Another Two-Stage Idea" to efficiently recognize suicide-related tweets. The results showed that tweets associated with suicide content were exclusively related to fear, sadness, and negative sentiments. The classification results showed the potential of emotions in facilitating the detection of suicide-related content online. Our findings offer valuable insights into ongoing research on the prevention of suicide risk and other mental-related disorders on Twitter. The proposed mechanism can contribute to the development of clinical decision support systems that deal with evidence-based guidelines and generate customized recommendations.

Research Data Related to this Submission

There are no linked research data sets for this submission. The following reason is given:
The data that has been used is confidential

*Highlights (for review)

Highlights

- The types of suicide- and non-suicide-related emotions on Twitter were examined.
- Suicide-related tweets were found to contain a higher proportion of fear, sadness, and negative emotions.
- The proposed approach was found to perform better than several existing approaches for classifying suicide tweets.

A Lexicon-based Approach to Detecting Suicide-related messages on Twitter

Abstract

Expression of emotion is an indicator that can contribute to the detection of mental health-related disorders. Suicide causes death to many people around the globe, and despite the suicide prevention strategies that have been employed over the years, only a few studies have explored the role of emotions in predicting suicidal behavior on social media platforms. This study explored the role of emotions from Twitter messages in detecting suicide-related content. We extracted and analyzed the characteristics of Twitter users' sentiment and behavior response (anger, fear, sadness, joy, positive, and negative) using NRC Affect Intensity Lexicon and SentiStrength techniques. A semi-supervised learning method was applied using the YATSI classifier or “Yet Another Two-Stage Idea” to efficiently recognize suicide-related tweets. The results showed that tweets associated with suicide content were exclusively related to fear, sadness, and negative sentiments. The classification results showed the potential of emotions in facilitating the detection of suicide-related content online. Our findings offer valuable insights into ongoing research on the prevention of suicide risk and other mental-related disorders on Twitter. The proposed mechanism can contribute to the development of clinical decision support systems that deal with evidence-based guidelines and generate customized recommendations.

Keywords: Twitter; suicidal thoughts; mental health; lexicon-based approach; semi-supervised learning; incidents detection

1. Introduction

A mental disorder is a medical condition characterized by signs or symptoms of a psychological (behavioral) nature. The prediction process of individuals’ mental health is usually achieved through the use of certain psychological concepts [1]. The literature showed that mental disorders can potentially contribute to suicide attempts in people [2, 3]. Suicidal thoughts affect an individual's emotional and cognitive state [4, 5]. However, due to the negative impact of suicide on society and public health [6, 7], understanding suicide-related behaviors in people has proven to be useful in developing effective prevention strategies [8-10]. This led many researchers (e.g., [11-13]) to consider exploring the content of social media platforms (e.g., Twitter, Facebook, Instagram, etc.) in an attempt to understand the negative experience of individuals who attempt suicide. These platforms allow users to express personal thoughts and emotions about general aspect of their life, typically in the form of messages [14]. By searching and downloading these messages, one can analyze the emotional reactions of a user prior to committing suicide [14]. In this context, Twitter, a popular social media platform, was recommended by many scholars (e.g., [12, 15]) as a reliable source of information to search and analyze people's mental health conditions through their messages (tweets).

Tweets are short messages of a maximum length of 280 characters and have the potential to speed up the dissemination of information in a social network. Due to the large number of tweets, machine learning algorithms are commonly used to extract, filter, and analyze the tweets in order to provide an automated recognition of suicide-related messages [16-18]. This has motivated scholars like Vioulès, Moulahi [19] to use machine learning algorithms for the identification of suicide-related risks from Twitter messages. The authors proposed an automatic recognition method to detect potential mental changes of Twitter users. Another work by Birjali, Beni-Hssane [20] predicted suicide incidents on Twitter using classification and sentiment analysis methods by constructing suicide vocabulary and recognition methods based on the WordNet lexicon. Varathan and Talib [14] proposed a new way to detect the suicide risk of online users with mental health problems based on data collected from Twitter, including users with a high risk of attempting suicide and others with previous suicide attempts. However, despite these efforts, previous studies have provided limited knowledge about how to use sentimental features (e.g., anger, fear, sadness, joy, positive, and negative) together with the classical features in the detection of suicide ideation in social media sites.

Our review of the literature also showed some key issues related to the detection of mental health problems from microblogs [21, 22]. There was a lack of accuracy in the detection of suicide-related incidents, which can be attributed to the lack of understanding the social conditions of suicide attempters [23]. Despite of the existing suicide prevention strategies, the rate of suicide has not changed significantly [24]. The currently used suicide detection methods have not managed to boost the recognition of suicide risk [25]. Based on these, this study proposed a new approach for categorizing suicide risk from social media posts. The proposed mechanism aims at extracting suicide-related emotions from Twitter messages using a semi-supervised learning technique. Semi-supervised learning is a popular type of machine learning that can be applied on small size labelled data and large unlabeled data sets [26]. Consequently, this study aimed to answer the following questions:

1. What are the types of suicide and non-suicide-related emotions being expressed on Twitter?
2. What is the role of these emotions in predicting suicide and non-suicide related messages?

To answer these questions, we performed sentiment analysis to extract emotions from suicide-related tweets. The performance of the utilized semi-supervised learning technique was examined with and without emotions. The contribution of sentimental features was observed throughout the suicide detection process.

2. Literature review

The private emotions of individuals are an effective means of investigating suicidal behavior. Our review of the literature showed that people who struggle in regulating their emotions are more vulnerable to suicide ideation in general [27]. To engage in a lethal suicide attempt, people must overcome their innate biological tendencies toward survival. The only way to do so is to regularly engage in avoidance behaviors, thus experiencing pain-related fear [28].

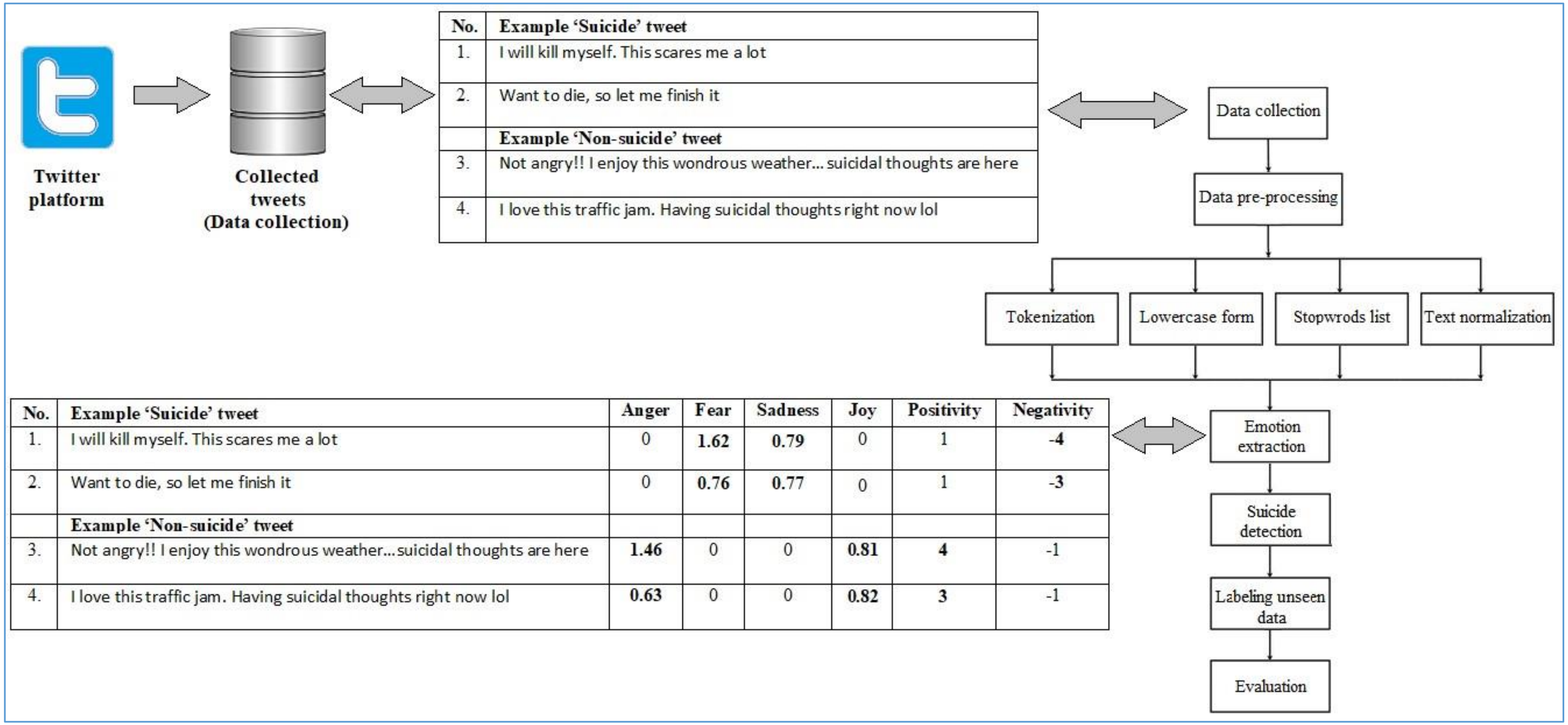
The dominant suicide theories are useful for providing knowledge on emotion dysregulation [29, 30], which can help in understanding suicidal behavior in people. Several theories in the literature have been adopted by researchers to explain suicidal actions from individual emotional expressions. For instance, the Cry of Pain model of suicidality has been widely used by previous studies to conceptualize suicidal behavior in individuals who have been regularly exposed to stressful situations [31]. This model presumes that individuals in such situations are more likely to experience negative emotions and low self-control [32]. One of the fundamental theories in sociology that has been widely used in many previous studies is Durkheim's theory[33]. This theory indirectly refers to the socioemotional theory of suicidal behavior, which is used to link both micro-level dynamics and the macro-level structural factors together. Based on this, one can argue that emotions can be effectively used to explain individuals' decisions in society. According to Abrutyn and Mueller [33], different social emotions can motivate a certain group of people to commit suicide. The authors looked at the potential association between emotions and cultural factors in explaining the individual's decision to commit suicide. In line with this, social media websites are a convenient medium for users to express their emotions and feelings regarding different matters [34, 35]. We can use these emotions to characterize psychiatric disorders with the help of machine learning techniques. Specifically, emotions embedded in social media messages can provide us with the necessary clues to predict suicide-related behaviors [14]. Therefore, studying emotions that are embedded in suicide messages along with sentiment analysis methods [36] may enable us to characterize various suicide-related criteria and recognize patients who authored such messages [5]. Burnap, Colombo [37] studied the potential of using machine learning algorithms in classifying suicide tweets. The authors utilized classifier distinguished emotional statements related to suicidal ideation and identified different suicide-related topics. They extracted several sets of features, including lexical, emotional, and idiosyncratic language features. Another study was conducted by Roberts, Roach [38] to examine a corpus that was obtained from Twitter labelled with seven types of emotions. The authors were mostly concerned about analyzing tweets and the distribution of emotions which are likely to appear in other corpora. They utilized an annotated corpus to train their machine- learning algorithm to automatically detect emotions in tweets. Based on this evidence, it can be said that emotions can play a significant role in people's decisions and action monitoring.

Other researchers like Luo, Du [12] proposed a method to deeply examine temporal patterns related to suicidal ideations on Twitter. Their proposed method carries a unique approach by detecting hidden suicide themes and modeling the average topic scores related to suicide risk. Another study conducted by Chiroma, Liu [39] performed text classification in the prediction of suicide on Twitter. The authors compared several machine learning classifiers (e.g., Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine (SVM)) to perform various suicide-ideation detection tasks. They reported that Decision Tree performed better than other classifiers in the detection of suicide-related communication. Astoveza, Obias [40] also used machine learning algorithms for predicting suicidal behavior from a set of tweets. They developed a predictive model using the Artificial Neural Network (ANN) to classify suicidal tweets that were obtained from performing an advanced search in Twitter. In addition, Du, Zhang [41] investigated the possibility of recognizing psychiatric stressors for suicide from Twitter using the Convolutional Neural Networks (CNN) classifier through a multiple-step pipeline (e.g., keyword- based retrieving, filtering, and classification). Sawhney, Manchanda [42] and Karamshuk, Shaw [43] proposed a computational technique for suicide identification of Twitter messages using a mixture of crowdsourcing and machine learning. Another study by Wang, Wan [44] explored the role of stylistic features in the detection of suicide ideation on Twitter using the SVM algorithm with a feature set of unigrams weighted by TF-IDF values. O'dea, Wan [15] explored the feasibility of performing an automated classification to predict the level of suicide risk severity among Twitter users by using a number of instructions and categories. Finally, a study by Abboute, Boudjeriou [10] explained the potential of using the 10-fold cross-validation technique in order to evaluate the performance of the Naïve Bayes classifier during the retrieval and detection of a specific suicidal risky behavior on Twitter.

From these results, it can be said that social media platforms like Twitter can provide a reliable source of information to help us identify and extract certain emotions. Consequently, using Twitter to analyze the embedded sentiment in suicide- related postswould contribute highly to the suicide detection process and lead to the establishment of a proper prevention strategy. Our review of the literature showed a lack of studies evaluating sentimental features in the detection of mental health disorders. Thus, the present study is the first attempt to examine the role of sentimental features in the detection of mental health disorders. It identified the types of emotions associated with suicide and non-suicide related statements using labelled and non-labelled data.

3. Method

This section explains the proposed method used in this study. The main stages of the development process are outlined in Figure 1. The data were collected with the use of legal tools designed by the Twitter developer account, where the necessary authentication tokens were obtained. Data on individual users were anonymized and used exclusively for the purposes of this study. Also, rigorous anonymity was maintained throughout the data processing and analysis phases. Here, we used the Waikato Environment for Knowledge Analysis (WEKA) open-source tool to analyze the extracted data.



No.	Example 'Suicide' tweet
1.	I will kill myself. This scares me a lot
2.	Want to die, so let me finish it
	Example 'Non-suicide' tweet
3.	Not angry!! I enjoy this wondrous weather... suicidal thoughts are here
4.	I love this traffic jam. Having suicidal thoughts right now lol

No.	Example 'Suicide' tweet	Anger	Fear	Sadness	Joy	Positivity	Negativity
1.	I will kill myself. This scares me a lot	0	**1.62**	**0.79**	0	1	**-4**
2.	Want to die, so let me finish it	0	**0.76**	**0.77**	0	1	**-3**
	Example 'Non-suicide' tweet						
3.	Not angry!! I enjoy this wondrous weather...suicidal thoughts are here	**1.46**	0	0	**0.81**	**4**	-1
4.	I love this traffic jam. Having suicidal thoughts right now lol	**0.63**	0	0	**0.82**	**3**	-1

Figure 1: Development stages

3.1 Data collection

We extracted two datasets from Twitter in order to assess the efficiency of the proposed method in the detection of suicide and non-suicide messages. The first dataset contains a total of 4,987 English tweets obtained from the Twitter Streaming API (Application Programming Interface) [45] using keywords like 'want to die', 'kill myself', 'suicide', 'suicidal thoughts', and 'contemplating suicide' based on the recommendations of Spates, Ye [46]. We invited two experts with 15 years of experience in mental health to help us identify and assess non-suicide tweets. The experts identified 500 tweets that were not discussing any aspects of suicidal behavior. For the second dataset, we used 500 tweets on suicidal ideation from sources recommended by Burnap, Colombo [37]. We combined the two datasets and built a training set of 1000 tweets (500 non-suicide and 500 suicide). Emotions embedded in these tweets were extracted using a lexicon-based method. Finally, a semi-supervised learning algorithm was applied to predict labels (suicide or non- suicide) of all the remaining 4487 (4,987 - 500) tweets. Finally, we compared the labelling of the 4487 tweets with manual labelling by the two experts to help us assess the accuracy of our approach. The following subsections discuss data pre-processing, emotion extraction, and classification in detail.

3.2 Data pre-processing

To prepare the training set (1000 tweets), we used the Bag-of-words model to extract the most relevant data features (e.g., words). This includes tokenizing these tweets and assigning weights to the resulting features. The weight of each word was set to either 0 or 1, appears or disappears, respectively. After that, all the words were converted into a lowercase form. We also used a Stopwords list technique followed by normalizing the length of each tweet using the L2 norm.

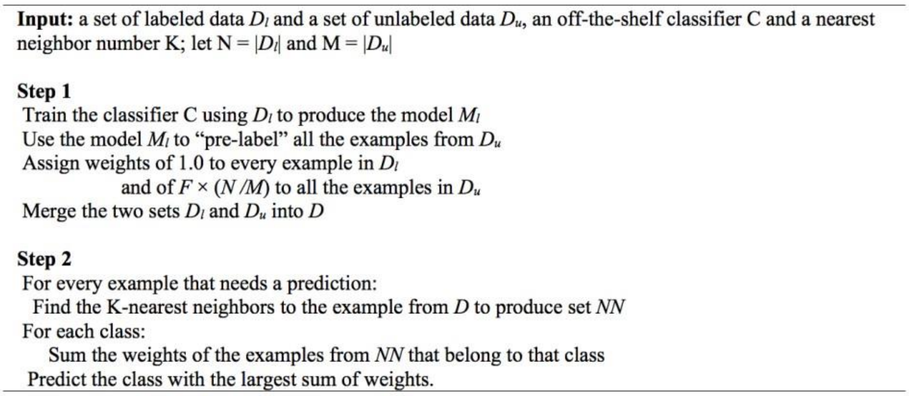
3.3 Emotion extraction

At this stage, users' sentimental features were extracted from their textual data using two main approaches: NRC Affect Intensity Lexicon [47] and SentiStrength [48, 49]. NRC Affect Intensity Lexicon allows extracting four popular emotions from the text: anger, fear, sadness, and joy. This was confirmed with the real-valued intensity for each emotion using the best-worst scaling approach. Specifically, a specific type of emotion was linked to a particular word using a score of either 0 (minimum amount of emotion) or 1 (maximum amount of emotion). We applied the NRC Affect Intensity Lexicon method independently on each of the two categories (non-suicide and suicide) in the training set (1000 tweets). Here, we calculated the most relevant data features by adding (counting) associations between words matching the given lexicons. In addition to the NRC Affect Intensity Lexicon approach, SentiStrength was also used to extract both positive and negative states for each tweet. SentiStrength was used to provide score values between '+1' (not positive) to '+5' (extremely positive) and '-1' (not negative) to '-5' (extremely negative).

The results obtained from the NRC Affect Intensity Lexicon phase showed that fear and sadness emotions were more frequent in suicide-related tweets than non-suicide texts (largely consisted of anger and joy emotions). In addition, the results from the SentiStrength tool revealed that suicidal tweets were largely associated with negative emotions as compared to the non-suicide set.

3.4 Suicide-ideation detection

We used semi-supervised learning technique at this stage to build a predictive model from the training set (1000 tweets), followed by labelling the other 3487 tweets. To find the best prediction model, two popular semi-supervised algorithms were applied. The first algorithm was the "Yet Another Two-Stage Idea" algorithm or "YATSI" [50]. This algorithm is based on the Random Forest model. We used YATSI its high capability in improving the predictive performance of the base classifier [50, 51] because of. It is a semi-supervised classification algorithm that uses both labeled and unlabeled data in two stages. At the first stage, a supervised classifier (random forest) is trained on the available training data. At the second stage, the model generated from the learning data is then used to pre-label all the test set instances. These pre-labelled instances are then used together with the original training data using a weighted nearest neighbour technique. The weights used by the nearest neighbour classifier are meant to help limit the level of trust of the algorithm during the labelling process of the model from the first step. The default value given to the weights of the training data was 1.0 while the weights of the pre-labelled test-data where N/M (N represents the number of training examples and M represents the number of test-examples). After adding the F parameter to the algorithm in order to regulate the weight of the test-examples to F * (N/M), we regulated the unlabeled data and the classifier built in the first stage. It is estimated that *F* values between 0.0 and 1.0 may reduce the influence on the test data and the learned model from the first stage. However, *F* values larger than 1.0 may increase the influence on the test data and the learned model. Below is the high-level pseudo-code for the two-stage YATSI algorithm according to Driessens, Reutemann [50]:



```
Input: a set of labeled data $D_l$ and a set of unlabeled data $D_u$, an off-the-shelf classifier C and a nearest
neighbor number K; let $N = |D_l|$ and $M = |D_u|$

Step 1
 Train the classifier C using $D_l$ to produce the model $M_l$
 Use the model $M_l$ to "pre-label" all the examples from $D_u$
 Assign weights of 1.0 to every example in $D_l$
             and of $F \times (N/M)$ to all the examples in $D_u$
 Merge the two sets $D_l$ and $D_u$ into $D$

Step 2
 For every example that needs a prediction:
   Find the K-nearest neighbors to the example from $D$ to produce set $NN$
 For each class:
     Sum the weights of the examples from $NN$ that belong to that class
  Predict the class with the largest sum of weights.
```

Along with the YATSI classifier, we used "Learning with Local and Global Consistency" or "LLGC" in order to reduce the number of instances in both labeled and unlabeled data while maintaining high precision [52]. LLGC is a graph-based approach used to solve the semi-supervised learning problem by satisfying both the local and global consistency assumptions. Due to the high performance of LLGC, it has been extended to clustering and ranking problems. According to Pfahringer, Leschi [53], LLGC works as follow:

1. Set up an affinity matrix *A*, where $A_{ij} =$ for $i \neq j$, and $A_{ii} = 0$.
2. Symmetrically normalize *A* yielding *S*, i.e. $S = D^{-0.5}AD^{-0.5}$ where *D* is a diagonal matrix with *D*(i, i) being the sum of the *i*-th row of *A*, which is also the sum of the *i*-th column of *A*, as *A* is a symmetrical matrix.

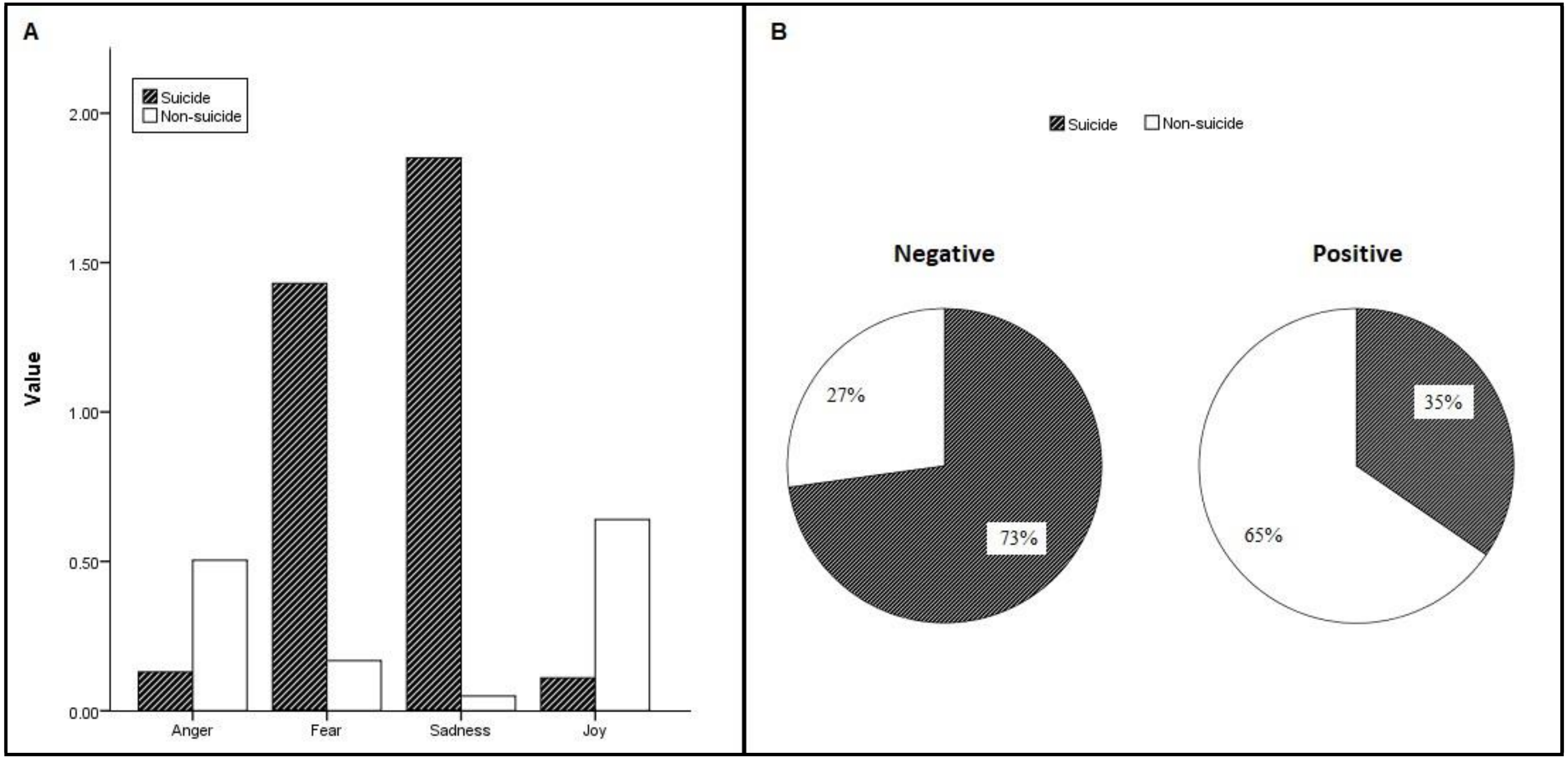
3. Setup matrix Y as a $n * k$ matrix, where n is the number of examples and k is the number of class values. Set $Y_{ik} = 1$, if the class value of example i is k. All other entries are zero, i.e. unlabeled examples are represented by all-zero rows.
4. Initialise $F(0) = Y$, i.e. start with the given labels.
5. Repeat $F(t + 1) = \alpha * S * F(t) + (1 - \alpha) * Y$ until F converges. α is a parameter to be specified by the user in the range [0, 1]. This iteration converges to: $F* = (1 - \alpha) * (I - \alpha * S)^{-1} * Y$

The values of class probability distributions are given to the normalized rows of $F*$ for every example. The main parameters for convergence are $0 \leq \alpha \leq 1$ holds, and that all eigenvalues of S are inside [−1, 1]. In this study, both YATSI and LLGC were used twice on the training set of two different features; unlike the second training set, the first training set contained no sentimental features (anger, fear, sadness, joy, positive, and negative). Following the recommendation of previous studies (e.g., [54-56]), the performance of the two algorithms was assessed using several evaluation metrics: Accuracy, Kappa statistic, Root Mean Squared Error (RMSE), Receiver Operating Characteristic (ROC), and Confusion matrix. The classification results showed that YATSI had the best classification performance. The results also demonstrated that the sentimental features boosted the performance of the classification process.

4. Results

4.1 Result of the sentiment analysis

The results of the NRC Affect Intensity Lexicon and SentiStrength are summarize in Figure 2. From Figure 2a, it can be observed that suicide-related tweets had higher level of fear (M = 1.43, SD = 0.15) and sadness emotions (M =1.85, SD = 0.09) than non-suicide tweets (fear: M = 0.16, SD = 0.02 and sadness: M = 0.06, SD = 0.21). In contrast, non-suicide tweets had higher mean values of anger (M = 0.50, SD = 0.34) and joy (M = 0.74, SD = 0.28) emotions than suicide-related tweets (anger: M = 0.13, SD = 0.05 and joy: M = 0.11, SD = 0.02). Nevertheless, to assess similarities/differences between the examined groups, a t-test was carried out and it revealed a significant difference (t = 2.88 $p<0.05$) between the two groups (suicide tweets had a higher level of fear and sadness emotions than non-suicide tweets). Non-suicide tweets had higher mean values of anger and joy emotions. Figure 2b shows the SentiStrength result which illustrate that the suicide tweets had a higher mean value of negative sentiment (M = -2, SD = 1.30) than non-suicide tweets (M = -1.81, SD = 1.16). The results also showed that non-suicide tweets had a positive sentiment score (M = 3.37, SD = 2.45) as compared to the suicide tweets (M = 1.78, SD = 1.14). In conclusion, suicide-related tweets had a higher mean value for fear, sadness, and negative emotions, whereas non-suicide tweets had a higher mean value for anger, joy, and positive emotions.



300 Figure 2: Result of NRC Affect Intensity Lexicon and SentiStrength

4.2 Result of the semi-supervised learning technique

As mentioned earlier, two semi-supervised algorithms (YATSI and LLGC) were used twice on 1000 tweets (with and without sentimental features). The prediction results are summarized in Table 1. In the first classification task (sentimental features), the YATSI algorithm had a higher classification accuracy (66.67 %) than the LLGC (59.38 %) method. As for the section classification task (with sentimental features), YATSI achieved a higher classification accuracy (86.97 %) than LLGC (75.42 %). In the first classification task, the YATSI classifier had a higher Kappa statistic value (11 %) than the LLGC algorithm (5 %). In the second classification task, YATSI had a higher Kappa statistic result (85 %) than LLGC (66 %). Furthermore, our results indicate that in the first classification task, the YATSI classifier had a lower RMSE value (42.13 %) than the LLGC algorithm (65.12 %). Also, in the second classification task, YATSI showed a lower RMSE value (2.05 %) than LLGC (25.84 %) (see Figure 3c). From the ROC curve result, it can be said that YATSI had a higher ROC value than LLGC (see Figure 3a, b).

Table 1: Classification result

Algorithm	Sentiment	Accuracy (%)	Kappa statistic (%)	RMSE (%)
YATSI	*Without*	66.67	11	42.13
	With	86.97	85	2.05
LLGC	*Without*	59.38	5	65.12
	With	75.42	66	25.84

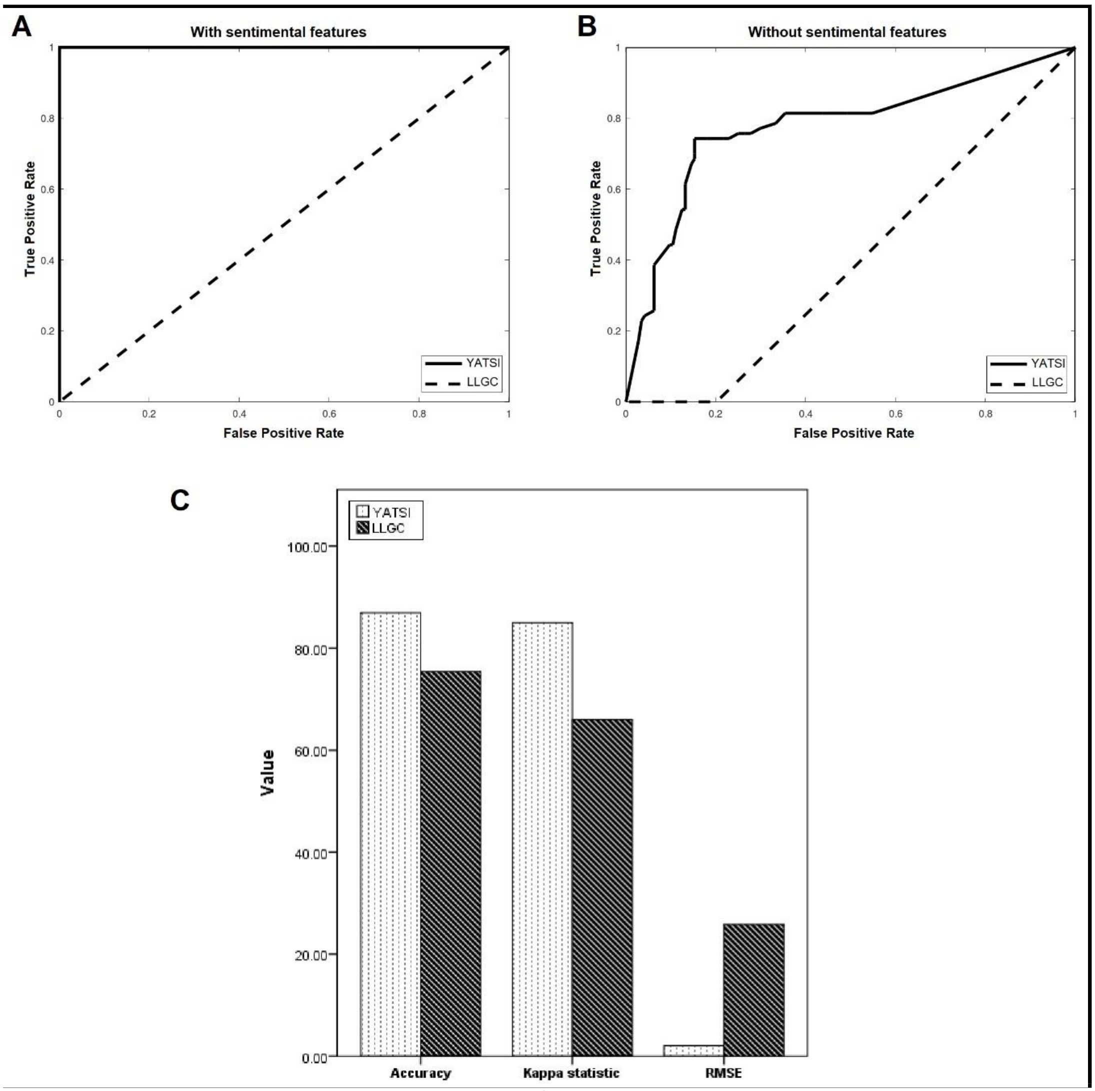


Figure 3: Evaluation metrics of the two algorithms

In this study, we generated the confusion matrix to further evaluate the classification performance of the examined classifiers. The confusion matrix is an essential method to analyze how well a classifier can predict instances (in this study, text) that belong to different classes. It measures the relationship between the predicted and actual instances; hence, for a classifier to show high performance, ideally most of the instances would be represented along the diagonal of the confusion matrix. The confusion matrix results revealed that the YATSI algorithm was able to recognize the instances for each class with a higher level of accuracy in both cases (with and without sentiment features) compared to the LLGC algorithm (see Figure 4). Based on these results, it can be concluded that the YATSI algorithm can be used efficiently to detect a suicide incident using sentimental features from microblogs. We used the YATSI classifier to label the unlabeled tweets as it had the best performance in the classification process. Then, to assess the labelling quality of the YATSI classifier, we compared the labelling of 4487 with the manual labelling of the two experts using Kappa statistic as the measure of agreement [55]. Kappa statistic results showed 94 % of agreement between the YATSI labelling and the manual labelling of the two experts (example tweets are highlighted in Table 2).

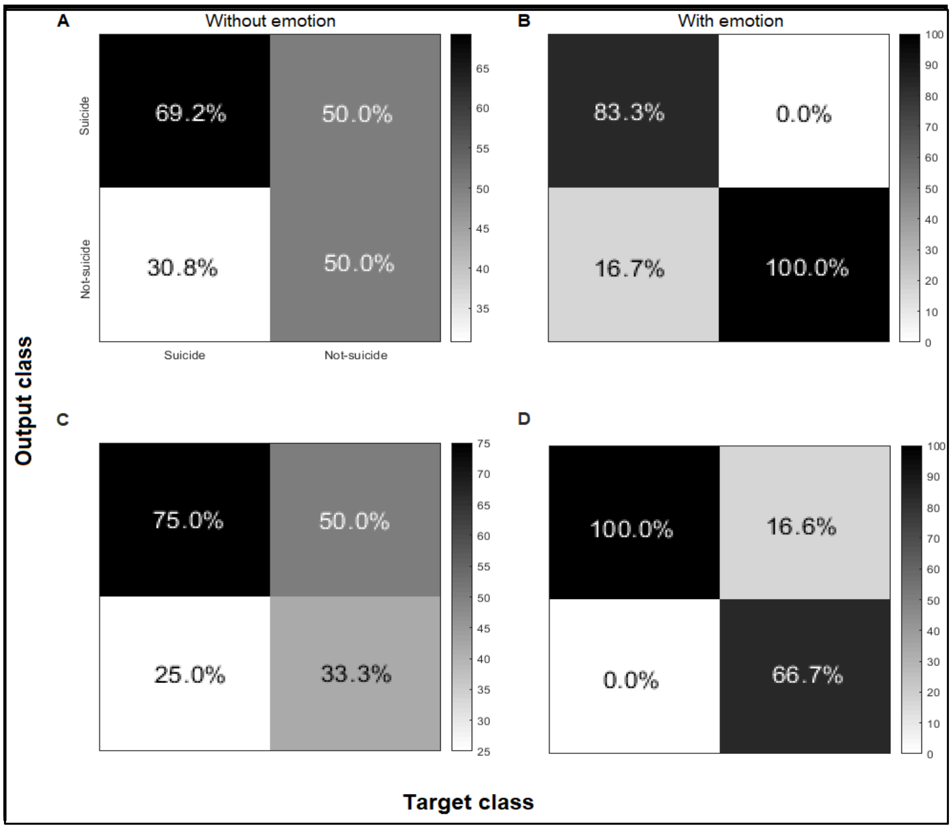


Figure 4: Confusion matrix

Table 2: Example of the classification results

No.	Statement	Predicted label
	True classification	
1.	To my friends, my work is done - Why wait?	Suicide
2.	I wrote jamb 7 times I didn’t kill myself	Non-suicide
3.	I am going to put myself to sleep now for a bit longer than usual	Suicide
4.	I can’t come and kill myself	Non-suicide
5.	They tried to get me - I got them first	Suicide
6.	I will not kill myself at this time	Non-suicide
	False classification	
1.	Life has become unbearable for me…Forgive me	Non-suicide
2.	I can’t make you happy	Suicide
3.	The future is just old age and illness and pain	Non-suicide
4.	I am so stressed, forgive me	Suicide
5.	I don’t want to hurt you or anybody so please forget about me	Non-suicide
6.	I will sleep the whole day	Suicide

4.3 A comparison of the proposed approach with previous studies

In order to evaluate the robustness of the proposed mechanism, we compared our results with relevant studies from the literature. The comparison result revealed that our method has a better performance than previous methods (see Table 3). This leads us to argue that having emotion extracted from NRC Affect Intensity Lexicon and SentiStrength can enhance the predictive capability of the collective classifier.

Table 3: Suicide detection techniques utilized in the literature

	Study	Approach	Result (%)
1.	Luo, Du [12]	CNN	83
2.	Vioulès, Moulahi [19]	Sequential Minimal Optimization	66.4
3.	Chiroma, Liu [39]	Decision tree	77.9
4.	Astoveza, Obias [40]	ANN--Multi-Layer Perceptron classifier	65
5.	Du, Zhang [41]	CNN	83
		Recurrent neural networks based psychiatric stressors recognition	53.25
6.	Sawhney, Manchanda [42]	Random Forest	85.8
7.	Burnap, Colombo [37]	Rotation Forest algorithm and a Maximum Probability voting classification decision method	72.8
8.	Karamshuk, Shaw [43]	Deep convolutional neural network architecture CharSCNN	71
9.	Wang, Wan [44]	Support vector machine (SVM)	66
10.	O'dea, Wan [15]	SVM	76
11.	Abboute, Boudjeriou [10]	Naïve Bayes via Leave One Out validation	63.15
		Naïve Bayes via 10-fold Cross-Validation	63.27
12.	The proposed method	YATSI	86.97

5. Discussions and implications

Our results showed that the type of emotions may vary between suicide and non-suicide related tweets. From the result, it can be observed that suicide-related statements/tweets were associated with higher rates of fear, sadness, and negative emotions than non-suicide tweets which reported higher rates of anger, joy, and positive emotions. We also found that the YATSI classifier had the best classification result when predicting suicide-related tweets. The emotion analysis result illustrated that having data with sentimental features can enhance the overall classification task. Our results extended the findings of previous research that examined emotions in suicide incidents. For example, a lack of positive emotions was found among those who attempted suicide. In addition, negative emotions were found to be highly correlated with suicidal actions [57].

Moreover, sentiments like sadness, hopelessness, anxiety, guilt, worthlessness, anger and irritability were those commonly observed in mental-health cases. Our analysis of users' sentiments, embedded in their suicide-related messages, revealed that sadness was the most frequent type of emotion [58]. Seidlitz, Conwell [59] observed that people who attempted suicide may experience a lower amount of positive sentiment than those who did not commit suicide. The authors also stated that anxiety could be an important feature in the detection process of suicide incidents. These findings are in line with our results, especially when taking into consideration the strength of the relationship between anxiety and fear emotions found in the literature [60]. In fact, both fear and sadness are correlated to each other. This can be explained by the fact that individuals who experience fear tend to express sadness and isolation, which may trigger rapid contemplation of suicide [33]. However, previous studies (e.g., [10, 18]) pointed out that fear and other emotions like anger and major aggressiveness may be found in suicide-related communication. Rogers, Kelliher- Rabon [61] stated that anger may indirectly be associated with suicide ideation/behavior, mainly because it is related to perceived burdensomeness and satisfaction with community. In other words, people who experience anger more often may tend to avoid communicating with other individuals, which in turn could lead to a lower level of social presence [62].

This study showed the extensive role of individuals' emotions in shaping their suicide-related behavior. This can be clearly shown by the presence of negative emotions in suicide statements. Based on this evidence, it can be said that our approach has the potential to characterize the suicide risk on social media platforms, which can be viewed as a step to prevent/reduce suicide incidents. The use of emotions in the recognition of suicidal behavior can help clinicians in monitoring patients' mental conditions via social media platforms. In addition, the proposed mechanism has the potential to contribute to the current clinical decision support system that implement a temporal risk profile of suicidal behaviors. This can help to develop interventions and find potential suicide victims at an early stage. Our mechanism can also be implemented to detect other psychiatric disorders on microblogs by characterizing the specific types of emotions that can be linked to illness behavior.

6. Limitations and future work

Despite the efficiency of the proposed method, some limitations still remain. For example, this study focused on English tweets in the analysis because it is the most popular language used in the world. We used specific keywords to search for suicide-related tweets; therefore, using other keywords may probably result in new features that would contribute to the overall detection of suicidal behavior. This study extracted certain sentimental features: anger, fear, sadness, joy, positive, and negative; thus, future work could explore other emotions and examine their relation to the suicide phenomenon. Finally, in this work, we studied suicide (a type of mental health disorder) due to the magnitude of this illness in our society. Future studies could apply our method on other mental illness and conduct further investigation of other types of emotions to understand other aspects of mental health disorder. In addition, the strongest predictor of suicidal thoughts and suicide is previous similar behaviour. This could be examined in the future by conducting longitudinal research. We also believe that future studies can focus on impulsivity as a major factor in going from suicidal thoughts to attempting suicide. Again, this can be identified in messages. It is the combination of previous suicidal behaviour, current suicidal-related emotions and impulsivity that leads to actually committing suicide.

7. Conclusion

This study proposed a novel method for detecting suicide-related risks on Twitter based on certain expressions of emotions (anger, fear, sadness, joy, positive, and negative) within tweets. We used a semi-supervised learning technique (the YATSI classifier) to predict suicide-related tweets. Our results revealed that suicide tweets had frequently higher rates of fear, sadness, and negative emotions than non-suicide tweets. We found that adding sentimental features into the detection of suicidal incidents can improve the performance of the classification process. These findings offer valuable insights into the prevention of suicidal risk and other mental health behaviors on microblogs. The proposed mechanism can potentially contribute to the development of clinical decision support systems that deals with suicidal behaviors.

References

1. Spitzer, R.L., J. Endicott, and J.-A.M. Franchi. Medical and mental disorder: Proposed definition and criteria. in *Annales Médico-psychologiques, revue psychiatrique*. 2018. Elsevier.
2. Hoertel, N., et al., Generalizability of clinical trial results for bipolar disorder to community samples: findings from the National Epidemiologic Survey on Alcohol and Related Conditions. The Journal of clinical psychiatry, 2013.
3. Luo, J., et al. Exploring Temporal Patterns of Suicidal Behavior on Twitter. in *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*. 2018. IEEE.
4. Cheung, G., S. Merry, and F. Sundram, Late-life suicide: Insight on motives and contributors derived from suicide notes. *Journal of affective disorders*, 2015. **185**: p. 17-23.
5. Desmet, B. and V. Hoste, Emotion detection in suicide notes. Expert Systems with Applications, 2013. **40**(16): p. 6351-6358.
6. Cerel, J., J.R. Jordan, and P.R. Duberstein, The impact of suicide on the family. *Crisis*, 2008. ***29***(1): p. 38-44.
7. Levine, H., Suicide and its impact on campus. *New Directions for Student Services*, 2008. ***2008***(121): p. 63-76.
8. Ho, T., et al., Suicide notes: what do they tell us? *Acta Psychiatrica Scandinavica*, 1998. ***98***(6): p. 467-473.
9. Zalsman, G., et al., Suicide prevention strategies revisited: 10-year systematic review. *The Lancet Psychiatry*, 2016. ***3***(7): p. 646-659.
10. Abboute, A., et al. Mining twitter for suicide prevention. in *International Conference on Applications of Natural Language to Data Bases/Information Systems*. 2014. Springer.
11. Sedgwick, R., et al., Social media, internet use and suicide attempts in adolescents. *Current opinion in psychiatry*, 2019. ***32***(6): p. 534.
12. Luo, J., et al., Exploring temporal suicidal behavior patterns on social media: Insight from Twitter analytics. Health informatics journal, 2019: p. 1460458219832043.

13. Lopez- Castroman, J., et al., Mining social networks to improve suicide prevention: A scoping review. *Journal of neuroscience research*, 2020. ***98***(4): p. 616-625.

14. Varathan, K.D. and N. Talib. Suicide detection system based on Twitter. in *2014 Science and Information Conference*. 2014. IEEE.

15. O'dea, B., et al., Detecting suicidality on Twitter. *Internet Interventions*, 2015. ***2***(2): p. 183-188.

16. Jashinsky, J., et al., Tracking suicide risk factors through Twitter in the US. Crisis, 2014.

17. Leiva, V. and A. Freire. Towards suicide prevention: early detection of depression on social media. in *International Conference on Internet Science*. 2017. Springer.

18. O'dea, B., et al., A linguistic analysis of suicide-related Twitter posts. Crisis: *The Journal of Crisis Intervention and Suicide Prevention*, 2017. ***38***(5): p. 319.

19. Vioulès, M.J., et al., Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, 2018. ***62***(1): p. 7: 1-7: 12.

20. Birjali, M., A. Beni-Hssane, and M. Erritali, Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, 2017. ***113***: p. 65-72.

21. Liu, X., et al., Proactive Suicide Prevention Online (PSPO): machine identification and crisis management for Chinese social media users with suicidal thoughts and behaviors. *Journal of medical Internet research*, 2019. ***21***(5): p. e11705.

22. Wang, X., et al., Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods: Content Analysis. *JMIR Medical Informatics*, 2020. ***8***(7): p. e17958.

23. Ribeiro, J., et al., Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychological medicine*, 2016. ***46***(2): p. 225-236.

24. Adamou, M., et al. Mining free-text medical notes for suicide risk assessment. in *Proceedings of the 10th hellenic conference on artificial intelligence*. 2018. ACM.

25. Walsh, C.G., J.D. Ribeiro, and J.C. Franklin, Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, 2017. ***5***(3): p. 457-469.

26. Keyvanpour, M.R. and M.B. Imani, Semi-supervised text categorization: Exploiting unlabeled data using ensemble learning algorithms. *Intelligent Data Analysis*, 2013. ***17***(3): p. 367-385.

27. Selby, E.A., et al., An exploration of the emotional cascade model in borderline personality disorder. *Journal of abnormal psychology*, 2009. ***118***(2): p. 375.

28. Law, K.C., L.R. Khazem, and M.D. Anestis, The role of emotion dysregulation in suicide as considered through the ideation to action framework. *Current Opinion in Psychology*, 2015. **3**: p. 30-35.

29. Klonsky, E.D., B.Y. Saffer, and C.J. Bryan, Ideation-to-action theories of suicide: a conceptual and empirical update. *Current Opinion in Psychology*, 2018. **22**: p. 38-43.

30. Gunn, J.F. and D. Lester, Theories of suicide: *Past, present and future*. 2015: Charles C Thomas Publisher.

31. Rasmussen, S.A., et al., Elaborating the cry of pain model of suicidality: Testing a psychological model in a sample of first- time and repeat self- harm patients. *British Journal of Clinical Psychology*, 2010. **49**(1): p. 15-30.

32. Hatkevich, C., F. Penner, and C. Sharp, Difficulties in emotion regulation and suicide ideation and attempt in adolescent inpatients. *Psychiatry research*, 2019. **271**: p. 230-238.

33. Abrutyn, S. and A.S. Mueller, Are suicidal behaviors contagious in adolescence? Using longitudinal data to examine suicide suggestion. *American Sociological Review*, 2014. **79**(2): p. 211-227.

34. Sun, X., et al., Detecting users' anomalous emotion using social media for business intelligence. *Journal of Computational Science*, 2018. **25**: p. 193-200.

35. Wu, P., et al., Social media opinion summarization using emotion cognition and convolutional neural networks. *International Journal of Information Management*, 2020. **51**: p. 101978.

36. Pestian, J.P., et al., Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 2012. **5**: p. BII. S9042.

37. Burnap, P., et al., Multi-class machine classification of suicide-related communication on Twitter. *Online social networks and media*, 2017. **2**: p. 32-44.

38. Roberts, K., et al. EmpaTweet: Annotating and Detecting Emotions on Twitter. In *Lrec*. 2012. Citeseer.

39. Chiroma, F., H. Liu, and M. Cocea. Text classification for suicide related tweets. in *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*. 2018. IEEE.

40. Astoveza, G., et al. Suicidal Behavior Detection on Twitter Using Neural Network. in *TENCON 2018-2018 IEEE Region 10 Conference*. 2018. IEEE.

41. Du, J., et al., Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making*, 2018. **18**(2): p. 43.

42. Sawhney, R., et al. A computational approach to feature extraction for identification of suicidal ideation in tweets. in *Proceedings of ACL 2018, Student Research Workshop*. 2018.

43. Karamshuk, D., et al., Bridging big data and qualitative methods in the social sciences: A case study of Twitter responses to high profile deaths by suicide. Online Social Networks and Media, 2017. **1**: p. 33-43.

44. Wang, Y., S. Wan, and C. Paris. The role of features and context on suicide ideation detection. in *Proceedings of the Australasian Language Technology Association Workshop 2016*. 2016.

45. Sarsam, S.M., H. Al-Samarraie, and B. Omar. Geo-spatial-based emotions: A mechanism for event detection in microblogs. in *Proceedings of the 2019 8th international conference on software and computer applications*. 2019.

46. Spates, K., X. Ye, and A. Johnson, "I just might kill myself": Suicide expressions on Twitter. *Death studies*, 2020. **44**(3): p. 189-194.

47. Mohammad, S.M., Word affect intensities. arXiv preprint arXiv:1704.08798, 2017.

48. Thelwall, M., The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength, in Cyberemotions. 2017, Springer. p. 119-134.

49. Culpeper, J., et al., Measuring emotional temperatures in Shakespeare's drama. English Text Construction, 2018. **11**(1): p. 10-37.

50. Driessens, K., et al. Using weighted nearest neighbor to benefit from unlabeled data. in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2006. Springer.

51. Imam, N., B. Issac, and S.M. Jacob, A Semi-Supervised Learning Approach for Tackling Twitter Spam Drift. *International Journal of Computational Intelligence and Applications*, 2019. ***18***(02): p. 1950010.

52. Zhou, D., et al. *Learning with local and global consistency*. In *Advances in neural information processing systems*. 2004.

53. Pfahringer, B., C. Leschi, and P. Reutemann. *Scaling up semi-supervised learning: An efficient and effective LLGC variant*. in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2007. Springer.

54. Sarsam, S.M., et al., A real-time biosurveillance mechanism for early-stage disease detection from microblogs: a case study of interconnection between emotional and climatic factors related to migraine disease. *NetMAHIB*, 2020. ***9***(1): p. 32.

55. Al-Samarraie, H., S.M. Sarsam, and H. Guesgen, *Predicting user preferences of environment design: A perceptual mechanism of user interface customisation. Behaviour & Information Technology*, 2016. ***35***(8): p. 644-653.

56. Sarsam, S.M. *Reinforcing the decision-making process in chemometrics: Feature selection and algorithm optimization*. In *Proceedings of the 2019 8th international conference on software and computer applications*. 2019.

57. Kiosses, D.N., K. Szanto, and G.S. Alexopoulos, Suicide in older adults: the role of emotions and cognition. *Current psychiatry reports*, 2014. ***16***(11): p. 495.

58. Pestian, J.P., P. Matykiewicz, and M. Linn-Gust, What's in a note: construction of a suicide note corpus. *Biomedical informatics insights*, 2012. ***5***: p. BII. S10213.

59. Seidlitz, L., et al., Emotion traits in older suicide attempters and non-attempters*.* *Journal of Affective Disorders*, 2001. ***66***(2-3): p. 123-131.

60. Turner, J.H., *Human emotions: A sociological theory*. 2007: Routledge.

61. Rogers, M.L., et al., Negative emotions in veterans relate to suicide risk through feelings of perceived burdensomeness and thwarted belongingness. *Journal of Affective Disorders*, 2017. ***208***: p. 15-21.

62. Hawkins, K.A., et al., An examination of the relationship between anger and suicide risk through the lens of the interpersonal theory of suicide. *Journal of Psychiatric Research*, 2014. ***50***: p. 59-65.