Hindawi
Complexity
Volume 2018, Article ID 6157249, 10 pages
https://doi.org/10.1155/2018/6157249

Research Article

Supervised Learning for Suicidal Ideation Detection in Online User Content

Shaoxiong Ji,[1,2] **Celina Ping Yu**,[3] **Sai-fu Fung**,[4] **Shirui Pan**,[5] **and Guodong Long**[5]

[1]*University of Queensland, Brisbane, Australia*
[2]*University of Technology Sydney, Sydney, Australia*
[3]*Global Business College of Australia, Melbourne, Australia*
[4]*City University of Hong Kong, Kowloon, Hong Kong*
[5]*Centre for Artificial Intelligence, University of Technology Sydney, Sydney, Australia*

Correspondence should be addressed to Shirui Pan; shirui.pan@uts.edu.au and Guodong Long; guodong.long@uts.edu.au

Received 1 February 2018; Revised 16 May 2018; Accepted 17 July 2018; Published 9 September 2018

Academic Editor: Gao Cong

Early detection and treatment are regarded as the most effective ways to prevent suicidal ideation and potential suicide attempts—two critical risk factors resulting in successful suicides. Online communication channels are becoming a new way for people to express their suicidal tendencies. This paper presents an approach to understand suicidal ideation through online user-generated content with the goal of early detection via supervised learning. Analysing users' language preferences and topic descriptions reveals rich knowledge that can be used as an early warning system for detecting suicidal tendencies. Suicidal individuals express strong negative feelings, anxiety, and hopelessness. Suicidal thoughts may involve family and friends. And topics they discuss cover both personal and social issues. To detect suicidal ideation, we extract several informative sets of features, including statistical, syntactic, linguistic, word embedding, and topic features, and we compare six classifiers, including four traditional supervised classifiers and two neural network models. An experimental study demonstrates the feasibility and practicability of the approach and provides benchmarks for the suicidal ideation detection on the active online platforms: Reddit SuicideWatch and Twitter.

1. Introduction

Suicide might be considered as one of the most serious social health problems in the modern society. Many factors can lead to suicide, for example, personal issues, such as hopelessness, severe anxiety, schizophrenia, alcoholism, or impulsivity; social factors, like social isolation and overexposure to deaths; or negative life events, including traumatic events, physical illness, affective disorders, and previous suicide attempts. Thousands of people around the world fall victims to suicide every year, making suicide prevention become a critical global public health mission.

Suicidal ideation or suicidal thoughts are people's thoughts of committing suicide. It can be regarded as a risk indicator of suicide. Suicidal thoughts include fleeting thoughts, extensive thoughts, detailed planning, role playing, and incomplete attempts. According to a WHO report [1], 788,000 people estimated worldwide committed suicide in 2015. And a large number of people, especially teenagers, were reported having suicidal ideation. Thus, one possible approach to preventing suicide effectively is early detection of suicidal ideation.

With the widespread emergence of mobile Internet technologies and online social networks, there is a growing tendency for people to talk about their suicide intentions in online communities. This online content could be helpful for detecting individuals' intentions and their suicidal ideation. Some people, especially adolescents, choose to post their suicidal thoughts in social networks, ask about how to commit suicide in online communities, and enter into online suicide pacts. The anonymity of online communication also allows people to freely express the pressures and anxiety they

suffer in the real world. This online user-generated content provides another possible angle for early suicide detection and prevention.

Previous research on suicide understanding and prevention mainly concentrates on its psychological and clinical aspects [2]. Recently, many studies have turned to natural language processing methods and classifying questionnaire results via supervised learning, which learns a mapping function from labelled training data [3]. Some of these researches have used the "International Personal Examination Screening Questionnaire," and analysed suicide blogs and posts from social networking websites. However, these studies have their limitations. (1) From both a psychological and a clinical perspective, collecting data and/or patients is typically expensive, and some online data may help in understanding thoughts and behaviours. (2) Simple feature sets and classification models are not predictive enough to detect suicidal tendencies.

In this paper, we investigate the problem of suicidal ideation detection in online social websites, with a focus on understanding and detecting the suicidal thoughts in online user content. We perform a thorough analysis of the content, the language preferences, and the topic descriptions to understand the suicidal thoughts from a data mining perspective. Six different sets of informative features were extracted and six supervised learning algorithms were compared to detect suicidal ideation within the data. It is a novel application of automatic suicidal intention detection on social content with the combination of our proposed effective feature engineering and classification models.

This paper makes notable contributions and novelties to the literature in the following respects:

(1) Knowledge discovery: this is a novel application of knowledge discovery and data mining to detect suicidal ideation in online user content. Previous work in this field has been conducted by psychological experts with statistical analysis; this approach reveals knowledge on suicidal ideation from a data analytic perspective. Insights from our analysis reveal that suicidal individuals often use personal pronouns to show their ego. They are more likely to use words expressing negativity, anxiety, and sadness in their dialogue. They are also more likely to choose the present tense to describe their suffering and the future tense to describe their hopelessness and plans for suicide.

(2) Dataset and platform: this paper introduces the Reddit platform and collects a new dataset for suicidal ideation detection. Reddit's SuicideWatch BBS is a new online channel for people with suicidal ideation to express their anxiety and pressures. Social volunteers respond in positive, supportive ways to relieve the depression and hopefully prevent potential suicides. This data source is not only useful for suicide detection but also for studying how to effectively prevent suicide through effective online communication.

(3) Features, models, and benchmarking: rather than using basic models with simple features for suicidal ideation detection, this approach (1) identifies informative features from a number of perspectives, including statistical, syntactic, linguistic, word embedding features, and topic features; (2) compares with different classifiers from both traditional and deep learning perspectives, such as support vector machine [4], Random Forest [5], gradient boost classification tree (GBDT) [6], XGBoost [7], multilayer feed forward neural net (MLFFNN) [8], and long short-term memory (LSTM) [9]; and (3) provides benchmarks for suicidal ideation detection on SuicideWatch on Reddit, one active online forum for communication about suicide.

This paper is organised as follows. In Section 2, we review the related works on suicide analysis and detection. We introduce the datasets in Section 3 along with data exploration and knowledge discovery. Section 4 describes the classification and feature extraction methods. Section 5 is the experimental study. We conclude this paper in Section 6.

2. Related Works

Suicide detection has drawn the attention of many researchers due to an increasing suicide rate in recent years. The reasons of suicide are complicated and attributed to a complex interaction of many factors [10]. The research techniques used to examine suicide also span many fields and methods. For example, clinical methods may examine resting-state heart rate [11] and event-related instigators [12]. Classical methods also include using questionnaires to assess the potential risk of suicide and applying clinician-patient interactions [13].

The goal of text-based suicide classification is to determine whether candidates, through their posts, have suicidal ideation. Such techniques include suicide-related keyword filtering [14, 15] and phrase filtering [16].

Machine learning methods especially supervised learning and natural language processing methods have also been applied in this field. The main features consist of N-gram features, knowledge-based features, syntactic features, context features, and class-specific features [17]. Besides, word embedding [18] and sentence embedding [19] are well applied. Models for cybersuicide detection include regression analysis [20], ANN [21], and CRF [22]. Okhapkina et al. built a dictionary of terms pertaining to suicidal content and introduced term frequency-inverse document frequency (TF-IDF) matrices for messages and a singular vector decomposition for matrices [23]. Mulholland and Quinn extracted vocabulary and syntactic features to build a classifier for suicidal and nonsuicidal lyricists [24]. Huang et al. built a psychological lexicon dictionary and used an SVM classifier to detect cybersuicide [25]. Chattopadhyay [8] proposed a mathematical model using Beck's suicide intent scale and applying multilayer feed-forward neural network to classify suicide intent. Pestian et al. [26] and Delgado-Gomez et al. [27] compared the performance of different multivariate techniques.

Table 1: Annotation rules and examples of social texts.

Categories	Rules	Examples
Suicide text	(i) Expressing suicidal thoughts	*I want to end my life tonight.*
	(ii) Including potential suicidal actions	*Yesterday, I tried to cut my wrist, but failed.*
Nonsuicide text	(i) Formally discussing suicide	*The global suicide rate is increasing.*
	(ii) Referring to other's suicide	*I am so sad to hear that Robin Williams ended his life.*
	(iii) Not relevant to suicide	*I love this TV show and watch every week.*

The relevant extant research can also be viewed according to the data source.

2.1. Questionnaires. Mental disorder scale criteria such as DSM-IV (https://www.psychiatry.org/psychiatrists/practice/dsm) and ICD-10 (http://apps.who.int/classifications/icd10/browse/2016/en), and the "International Personal Disorder Examination Screening Questionnaire" (IPDE-SQ) provide good tools for evaluating an individual's mental status and their potential for suicide. Delgado-Gomez et al. classified the results of IPDE-SQs based on "Barrat's Impulsiveness Scale" (version 11) [28] and the "Holmes-Rahe Social Readjustment Rating Scale" to identify people likely to attempt suicide [27].

2.2. Suicide Notes. Suicide notes provide material for natural language processing. Previous approaches have examined suicide notes using content analysis [26], sentiment analysis [17, 29], and emotion detection [22]. In the age of cyberspace, suicide notes are now also written in the form of web blogs and can be identified as carrying the potential risk of suicide [14].

2.3. Online User Content. Cash et al. [30], Shepherd et al. [31], and Jashinsky et al. [16] have conducted psychology-based data analysis for content that suggests suicidal tendencies in the MySpace and Twitter social networks. Ren et al. explored accumulated emotional information from online suicide blogs [32]. O'Dea et al. developed automatic suicide detection on Twitter by applying logistic regression and SVM on TF-IDF features [33]. Reddit has also attracted much research interest. Huang and Bashir applied linguistic cues to analyse the reply bias [34]. De Choudhury et al. did many works on suicide-related topics in Reddit including the effect of celebrity suicides on suicide-related content [35] and the transition from mental health illness to suicidal ideation [36].

A questionnaire is a useful tool for collecting data, but it costs highly. Suicide notes are useful materials for training a classifier. The current dataset of suicide notes is quite small. Automatic detection on online user content will be a promising way for suicide detection and prevention. Our proposed method investigated a better solution with effective feature engineering on a bigger social dataset than the previous work. And it can adapt to real-world application with the ability of automatic detection compared with questionnaires.

3. Data and Knowledge

We collect the suicidal ideation texts from Reddit and Twitter and manually check all the posts to ensure they were correctly labelled. Our annotation rules and examples of posts appear in Table 1.

Table 2: Two balanced Reddit datasets.

Dataset	Subreddits
1	SuicideWatch versus others (nonsuicide)
2	SuicideWatch versus gaming
	SuicideWatch versus jokes
	SuicideWatch versus books
	SuicideWatch versus movies
	SuicideWatch versus AskReddit

3.1. Reddit Dataset. Reddit is a registered online community that aggregates social news and online discussions. It consists of many topic categories, and each area of interest within a topic is called a subreddit.

In this dataset, online user content includes a title and a body of text. To preserve privacy, we replace personal information with a unique ID to identify each user. We collected posts with potential suicide intentions from a subreddit called "SuicideWatch"(SW) (https://www.reddit.com/r/SuicideWatch/). Posts without suicidal content were sourced from other popular subreddits (https://www.reddit.com/r/all/, https://www.reddit.com/r/popular/). The collection of nonsuicidal data is totally a user-generated content, and the posts of news aggregation and administrator are excluded. To facilitate the study and demonstration, we will study the balanced dataset in Reddit and study imbalanced dataset in Twitter in the following subsection.

The Reddit dataset includes 3549 suicidal ideation samples and a number of nonsuicide texts. In particular, we construct two datasets for Reddit as shown in Table 2. The first dataset includes two subreddits in which one is from SuicideWatch and another is from popular posts in Reddit. The second dataset is composed of six subreddits that include SuicideWatch and another five hot topics: gaming (https://www.reddit.com/r/gaming/), jokes (https://www.reddit.com/r/Jokes/), books (https://www.reddit.com/r/books/), movies (https://www.reddit.com/r/movies/), and AskReddit (https://www.reddit.com/r/AskReddit/). In the second dataset, the combination of SuicideWatch with any other subreddit will be a new balanced subdataset, for example, suicide versus gaming and suicide versus jokes. These two datasets will be studied on Subsections 5.1 and 5.2 separately.

(a) Reddit word cloud

(b) Twitter word cloud

Figure 1: Word cloud visualisation of suicidal texts in Reddit and Twitter.

Table 3: Linguistic statistical information extracted by LIWC.

Average word count	Suicide	Nonsuicide
Personal nouns	30.01	14.6
Quantifiers	3.78	3.37
Positive emotion	5.61	7.84
Negative emotion	11.12	4.89
Anxiety	1.46	0.55
Sadness	3.86	0.63
Past focus	6.78	6.27
Present focus	34.81	17.86
Future focus	4.06	1.76
Family	1.07	0.82
Friend	1.02	0.78
Female references	0.95	1.35
Male references	1.03	2.40
Work	2.50	3.92
Money	0.60	1.38
Death	4.81	0.61
Swear words	1.47	1.62

3.2. Twitter Dataset. Many online users also want to talk about the suicidal ideation in social networks. However, Twitter is quite different with Reddit as (1) each tweet's length is limited in 140 characters (this limit is now 280 characters), (2) tweet users may have some social network friends from the real world while Reddit users are fully anonymous, and (3) the communication and interaction type are totally different between social networking websites and online forums.

The Twitter dataset is collected using a keyword filtering technique. Suicidal words and phrases include "suicide," "die," and "end my life." Many of the collected tweets have the suicidal-related words, but they possibly talk about a suicide movie or advertisement which does not contain suicidal ideation. Therefore, we manually checked and labeled collected tweets according to the annotation rules in Table 1. Finally, the Twitter dataset has totally 10,288 tweets with 594 tweets (around 6%) with suicidal ideation. This dataset is an imbalanced dataset and will be studied in Section 5.3.

3.3. Data Exploration and Knowledge Discovering. To understand suicidal individuals, we analysed the words, languages, and topics in online user content.

3.3.1. Word Cloud. Word clouds were used to provide a visual understanding of the data. The users' posts in Reddit and tweets in Twitter with potential suicide risk are showed separately in Figures 1(a) and 1(b). As we can see, suicidal posts frequently use words such as "life," "suicide," and "kill," providing a direct indication of the users' suicidal thoughts. Words expressing feelings or intentions are also frequently used, such as "feel," "want," and "know." For example, some suicidal posts wrote, "I feel like I have no one left and I want to end it," "I want to end my life," and "I don't know how much of it was psychological trauma."

In addition, the dominant words in these two social platforms have different styles due to the posting rules of the platforms. The Reddit users are willing to compose their posts in a specific way. For instance, they describe their life events and their stories about their friends. While the content in Twitter is much more straightforward with expressions like "want kill," "going kill," and "wanna kill." The details are usually not included in their tweets.

3.3.2. Language Preferences. Language preferences provide an overview of the statistical linguistic information of the data. The listed variables shown in Table 3 were extracted using LIWC 2015 [37]. All these categories are features based on word counts. We calculated the average value of each variable in both suicide-related texts and suicide-free posts. As shown in the table, content with or without suicidality quite differs in many items.

(i) Users with suicidal ideation use many personal pronouns to show their ego. For example, "I want to end my life."

(ii) They express more negative emotions, like anxiety and sadness. For example, "I was drowning in guilt and depression for several years after."

(iii) As for the tense, texts with suicidal ideation tend to use the present and future tense. They tend to use the present tense to describe their suffering, pain, and depression. For example, "I'm feeling so bad." The future tense is used to describe their hopeless feelings about the future and their suicide intentions. For example, "I'm eventually going to kill myself."

(iv) Both types of posts discuss family and friends and make female or male references.

Table 4: Topic words extracted from posts containing suicidal thoughts.

Number	Top 10 words for each suicide-related topics in SuicideWatch
1	Money, working, suicide, gun, fucked, come, yet, failed, erase, thats
2	Said, got, went, started, friend, back, father, told, mother, girl
3	Im, school, go, year, time, know, one, ive, day, got
4	Mm, dont, its, ive, cant, get, know, around, time, pain
5	Im, feel, like, want, know, friend, would, life, get, time
6	Imagine, cellophane, abandoned, anyone, medical, cheated, mr, surgery, yelling, letter
7	Im, want, life, like, get, feel, ive, know, year, even
8	Fucking, very, tomorrow, bottom, accept, sharp, n't, went, wife, attacked
9	Condition, suicide, also, hope, tx, california, chronic, jumping, crisis, age
10	Please, find, mother, car, social, live, need, accident, debt, month

(v) Unsurprisingly, more words related to death appear in texts about suicide. For example, "kill," "die," "end life," and "suicide."

(vi) Both types of posts contain a similar number of swear words.

One of the findings from Table 3 and Figure 1 is that people with suicidal thoughts tend to directly show their intentions in anonymous online communities when faced with some kinds of problem in the real world. Their posts often show negative feelings with strong ego and intention.

3.3.3. Topic Description. We extracted 10 topics from posts containing suicidal ideation using the latent Dirichlet allocation (LDA) [38] topic modelling method, as shown in Table 4. There are some Internet slangs such as "tx" (thanks) and abbreviations like "im" (I am) and "n't" ("negatory"). In the field of standard natural language processing, personal words like "I," "me," and "you" are stop words and should be removed, but we kept them in this exploration because they contain important information. Thus, there are many personal pronouns included in these topic words, which are identical to the results in Table 3.

Interestingly, we observed that posts containing suicidal themes could be summarised into three categories: internal factors, external social factors, and mixed internal/external factors. Specifically, internal factors, including words like "know" (topics 3, 4, 5, and 7), "want," "feel" and "like" (topics 5 and 7), and "hope" (topic 9) express people's feelings, intentions, and desires, while other words such as "money" and "working" (topic 1), "friend" (topics 2 and 5), "school" (topic 3), "surgery" (topic 6), "crisis" (topic 9), and "accident" (topic 10) indicate that posts are linked to social factors. In topic 3, 5, 9, and 10, both factors are represented.

4. Methods and Technical Solutions

4.1. Feature Processing. By preprocessing and cleaning the data in advance, we extracted several features including statistics, word-based features (e.g., suicidal words and pronouns), TF-IDF, semantics, and syntactics. Additionally, we used distributed features by training neural networks to embed word into vector representations, along with topic features extracted by LDA [38] as unsupervised features.

4.1.1. Statistical Features. User-generated posts are varied in length, and some statistical features can be extracted from texts. Some posts use short and simple sentences, while others use complex sentences and long paragraphs.

After segmentation and tokenisation, we captured statistical features as follows:

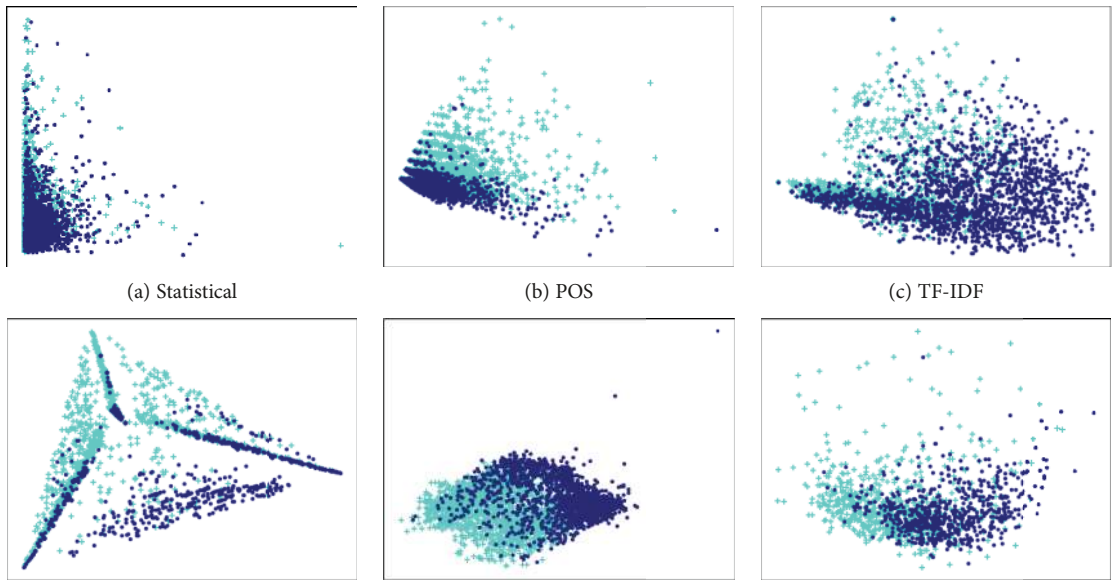
(i) The number of words, tokens, and characters in the title

(ii) The number of words, tokens, characters, sentences, and paragraphs in the text body

4.1.2. Syntactic Features: POS. Syntactic features are useful information in natural language processing tasks. We extracted parts of speech (POS) [39] as features for our suicidal ideation detection model to capture the similar grammatical properties in users' posts.

Common POS tags include nouns, verbs, participles, articles, pronouns, adverbs, and conjunctions. POS subgroups were also identified to provide more detail about the grammatical properties of the posts. Each post was parsed and tagged, and the number of each category in the title and text body was simply counted.

4.1.3. Linguistic Features: LIWC. Online users' posts usually contain emotions, relativity, and harassment words. Lexicons are widely applied for extracting these features. To analyse the linguistic and emotional features in the data, we used Linguistic Inquiry and Word Count [37] (LIWC 2015 (http://liwc.wpengine.com/)) which was proposed and developed by the University of Texas at Austin. This approach was used in a previous study [34]. The tool contains a powerful internally built dictionary for matching the target words in posts when parsing data. About 90 variables were output. In addition to word count-based features, it could extract features based on emotional tone, cognitive processes, perceptual processes, and many types of abusive words. Specific categories include word count, summary language, general descriptors, linguistic



(a) Statistical (b) POS (c) TF-IDF

(d) Topics (e) LIWC (f) Word embedding

Figure 2: Visualisation of extracted features using PCA.

dimensions, psychological constructs, personal concern, informal language markers, and punctuation.

4.1.4. Word Frequency Features: TF-IDF. Many kinds of expression are related to suicide. We used TF-IDF to extract these features and measure the importance of various words from both suicidal posts and nonsuicidal posts. TF-IDF measures the number of times that each word occurs in the documents and adds a penalty depending on the frequency of the word in the entire corpus.

4.1.5. Word Embedding Features. The distributed representation, which is able to preserve the semantic information in texts, is popular and useful for many natural language processing tasks. It embeds words into a vector space. There are several techniques for word embedding. We employed the *word2vec* ([18], https://code.google.com/archive/p/word2vec/) to derive a distributed semantic representation of the words.

There are two architectures for word2vec word embedding, that is, CBOW and Skip-gram. CBOW predicts the present word based on the context, Skip-gram predicts the closest words to the current word provided.

4.1.6. Topic Features. Suicidal posts and nonsuicidal posts talk about different topics which can provide good understanding for two categories. We applied the latent Dirichlet allocation (LDA) [38] to reveal latent topics in user posts. Each topic is a mixture probability of word occurrence in the topic, and each post is a mixture probability of topics.

Given the set of documents and the number of topics, we used LDA to extract the topics from each posts, then calculate the probability that each post belonged to every generated topics. Hence, the posts are represented by their thematic properties as probability vectors at the length of the number of topics.

(1) Feature Visualisation. To understand the informativeness of these feature sets, we visualise the features on the Reddit dataset in a 2-dimensional space by using principal component analysis (PCA) [40] in Figure 2. The results demonstrate that we indeed extract features that can largely separate the points in different classes. We will further validate the effectiveness of our feature sets in Section 5.

4.2. Classification Models. Suicidality detection in social content is a typical classification problem of supervised learning. Given a dataset $\{x_i, y_i\}_i^n$ consisting a set of texts $\{x_i\}_i^n$ with labels $\{y_i\}_i^n$, we trained a supervised classification model to learn the function from the training data pairs of input objects and supervisory signals:

$$y_i = F(x_i), \tag{1}$$

where $y_i = 1$ means that the expression x_i is "suicide text" (ST), otherwise $y_i = 0$ means "not suicide text (non-ST)." The training or learning of the classification model is to minimise the prediction error in the given training data. The prediction error is to be presented as a loss function $L(y, F(x))$ where y is the real label and $F(x)$ is the predicted label by using classification model. In summary, the goal of training algorithm is to obtain an optimal prediction model $F(x)$ by solving below optimisation task:

$$\widehat{F} = \underset{F}{\operatorname{argmin}} \mathbb{E}_{x,y}[L(y, F(x))]. \tag{2}$$

Different classification methods may have different definition of loss function and predefined structure of model. We employed both classical supervised learning classification

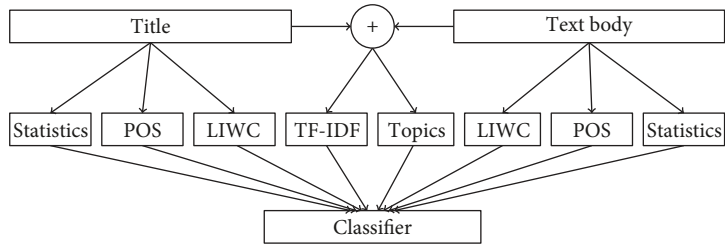


Figure 3: The model's structure for Reddit dataset.

methods and deep learning methods to solve the suicidal ideation classification task.

The structure of our feature extraction method is shown in Figure 3. As mentioned in Section 4.1, features comprised statistics, POS counts, LIWC features, TF-IDF vectors, and topic probability features. Among these features, we applied POS features and LIWC features to both the title and text body of user posts. We combined the title and the body into one piece of text to extract topic probability vectors and TF-IDF vectors. All extracted features were input to the classifiers.

5. Empirical Evaluation

5.1. Comparison and Analysis on Suicide versus Nonsuicide. This section compares various classification methods using different combinations of features with 10-fold cross validation (Our codes are available in https://github.com/shaoxiongji/sw-detection). The specific classification models include support vector machine [4], Random Forest [5], gradient boost classification tree (GBDT) [6], XGBoost [7], and multilayer feed-forward neural net (MLFFNN) [8]. SVM is able to solve problems that are not linearly separable in lower space by constructing a hyperplane in high-dimensional space. It can be adapted to many kinds of classification tasks [41, 42]. Random Forest, GBDT, and XGBoost are tree ensemble methods that use decision trees as base classifiers and produce a form of committee to gain better performance than any single base classifier. MLFFNN takes the different features as input and learns the combination of them with nonlinearity.

For comparison and to solve the problem of understanding the semantic meaning and syntactic structure of sentences, deep learning provides powerful performance [43]. We used long short-term memory (LSTM) [9] network, one state-of-the-art deep neural network. LSTM takes the title and text body of user posts with word embedding as its inputs and uses memory cell to preserve the state over long periods, capturing the long-term dependencies in long conversation detection.

As shown in Table 5, all methods' performance increases by combining more features on the whole. This observation validates the effectiveness and informativeness of our extracted features. However, the contribution each feature makes varies, which leads to fluctuations in the results of individual methods. The XGBoost had the best performance of the six methods when taking all groups of features as inputs. Although LSTM does not require feature processing and is renowned for its state-of-the-art performance in many other natural language processing tasks, it did not perform as well as some of the other ensemble learning methods with sufficient features in this case. Random Forest, GBDT, XGBoost, and MLFFNN with proper features produced better accuracy and F1 scores than LSTM on our Reddit dataset. Admittedly, deep learning with word embedding is rather convenient and typically achieves adequate results, even without complicated feature engineering.

The AUC performance measurement in each classification is the area under the receiver operating characteristic curve with all extracted features. In the last column of Table 5, the AUC has an increasing tendency with more combined features. The XGBoost method gains the highest AUC of 0.9569 while other methods have very similar AUC value above 0.9.

5.2. Suicide versus Single Subreddit Topics. To evaluate the classification on suicide with any other specific online communities, we extended our datasets and experiments to other specific subreddits, including "gaming," "jokes," "books," "movies," and "AskReddit."

The results are shown in Figure 4. Using the features extracted with our approach was a very effective way of classifying the suicidal ideation posts from another subreddit domain. In fact, the classification results on suicidal dataset versus the subreddit dataset were better than suicidal versus nonsuicidal dataset where the nonsuicidal samples are composed of multiple popular subreddit domains. In these experiments, XGBoost produced the best results on "movies" and "AskReddit" in terms of accuracy and F1 scores. LSTM and Random Forest outperformed the other models in "gaming" and "books," respectively.

5.3. Experiments on Twitter Dataset. To evaluate the performance of our proceeded features and the classification models, we do another experiment on our Twitter dataset. Tweet text without long text body is different with Reddit text. Thus, for the experimental setting, there is a slight difference between them. We exclude the number of paragraphs in statistical features, POS, and LIWC features of text bodies. The rest of the settings are similar to our previous experiment. Considering the class imbalance in Twitter data, we adopt undersampling techniques. The results are the average metrics of each undersampled data shown in Table 6. The receiver operating characteristic curves of

Table 5: Comparison of different methods using different features.

Methods	Features	Acc.	Prec.	Recall	F1-score	AUC
SVM	Statistics	0.8064	0.8045	0.8189	0.8116	0.8061
	Statistics + topic	0.8609	0.881	0.8406	0.8603	0.8613
	Statistics + topic + TF-IDF	0.8571	0.8414	0.8865	0.8634	0.8565
	Statistics + topic + TF-IDF + POS	0.8674	0.8545	0.8916	0.8727	0.8670
	Statistics + topic + TF-IDF + POS + LIWC	0.9123	0.9144	0.9133	0.9138	0.9123
Random Forest	Statistics	0.7732	0.8094	0.7258	0.7653	0.7741
	Statistics + topic	0.8973	0.8922	0.9082	0.9001	0.8971
	Statistics + topic + TF-IDF	0.8915	0.8795	0.912	0.8954	0.8911
	Statistics + topic + TF-IDF + POS	0.8986	0.8801	0.9273	0.9031	0.8981
	Statistics + topic + TF-IDF + POS + LIWC	0.9357	0.9213	0.9554	0.938	0.9353
GBDT	Statistics	0.7505	0.7632	0.7398	0.7513	0.7507
	Statistics + topic	0.898	0.8856	0.9184	0.9017	0.8976
	Statistics + topics + TF-IDF	0.896	0.89	0.9082	0.899	0.8958
	Statistics + topic + TF-IDF + POS	0.8928	0.8893	0.9018	0.8955	0.8926
	Statistics + topic + TF-IDF + POS + LIWC	0.9461	0.9354	0.9605	0.9478	0.9458
XGBoost	Statistics	0.7667	0.7822	0.7513	0.7664	0.7670
	Statistics + topic	0.8999	0.8938	0.912	0.9028	0.8997
	Statistics + topic + TF-IDF	0.9019	0.8941	0.9158	0.9049	0.9016
	Statistics + topic + TF-IDF + POS	0.9103	0.8998	0.9273	0.9133	0.9100
	Statistics + topic + TF-IDF + POS + LIWC	**0.9571**	0.9499	**0.9668**	**0.9583**	**0.9569**
MLFFNN	Statistics	0.7647	0.7742	0.7742	0.7742	0.7731
	Statistics + topic	0.8821	0.8740	0.8525	0.8631	0.8961
	Statistics + topic + TF-IDF	0.8606	0.8369	0.8401	0.8385	0.8855
	Statistics + topic + TF-IDF + POS	0.9068	0.9038	0.8868	0.8952	0.9369
	Statistics + topic + TF-IDF + POS + LIWC	0.9283	0.9391	0.9205	0.9295	0.9403
LSTM	word2vec word embedding	0.9266	**0.9786**	0.8750	0.9239	0.9276

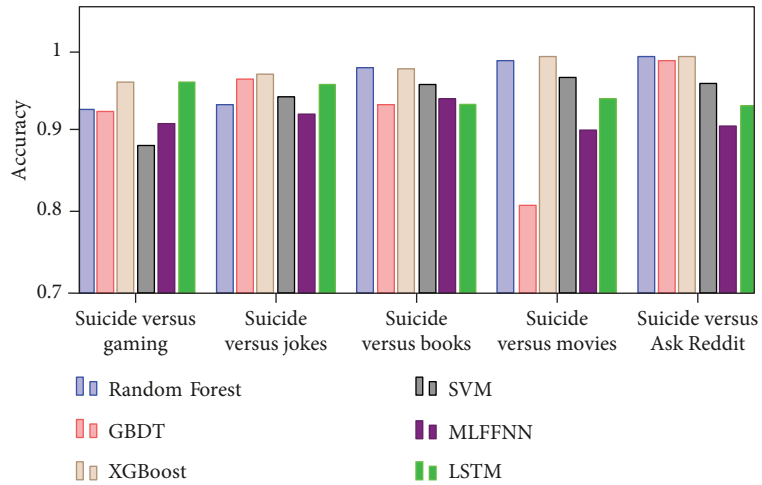


Figure 4: Classification for suicidal ideation of SuicideWatch versus other six subreddits.

these methods are showed in Figure 5. In these dataset, Random Forest gains better performance than most models except for the metric of precision in which the MLFFNN gains a slightly better result.

6. Conclusion

The amount of text keeps growing with the popularisation of social networking services. And suicide prevention

Table 6: Comparison of different models using all processed features on Twitter data.

Model	Acc.	Prec.	Recall	F1	AUC
Random Forest	**0.9638**	0.9638	**0.9917**	**0.9646**	**0.9862**
GBDT	0.9500	0.9413	0.9603	0.9503	0.9825
XGBoost	0.9591	0.9425	0.9782	0.9597	0.9843
SVM	0.9485	0.9261	0.9755	0.9497	0.9813
MLFFNN	0.9412	**0.9661**	0.9194	0.9421	0.9823
LSTM	0.9108	0.9399	0.8802	0.9059	0.9747

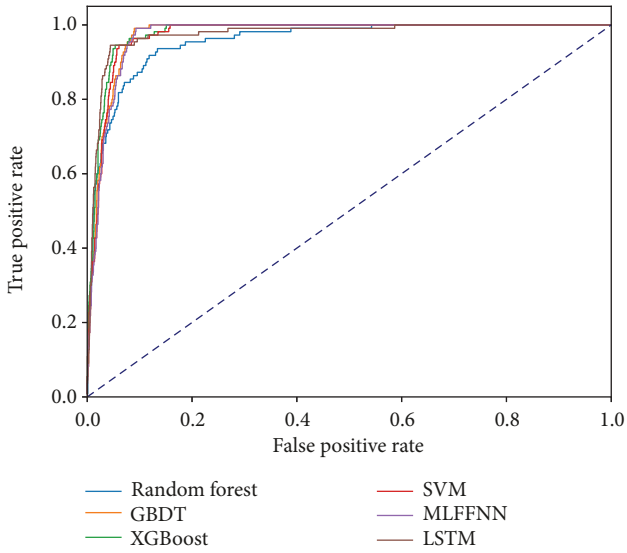


Figure 5: The receiver operating characteristic curve of six methods with all processed features.

remains an important task in our modern society. It is therefore essential to develop new methods to detect online texts containing suicidal ideation in the hope that suicide can be prevented.

In this paper, we investigated the problem of suicidality detection in online user-generated content. We argue that most work in this field was conducted by psychological experts with statistical analysis, which is limited by the cost and privacy issue in obtaining data. By collecting and analysing the anonymous online data from an active Reddit platform and Twitter, we provide rich knowledge that can complement the understanding of suicidal ideation and behaviour. Though applying feature processing and classification methods to our carefully built datasets, Reddit and Twitter, we evaluated, analysed, and demonstrated that our framework can achieve high performance (accuracy) in distinguishing suicidal thoughts out of normal posts in online user content.

While exploiting more effective feature sets, complex models or other factors such as temporal information may improve the detection of suicidal ideation—these will be our future directions; the contribution and impact of this paper are threefold: (1) delivering rich knowledge in understanding suicidal ideation, (2) introducing datasets for the research community to study this significant problem, and (3) proposing informative features and effective models for suicidal ideation detection.

Data Availability

The data used to support the findings of this study are available from the first author upon request (email: shaoxiong.ji@uq.edu.au).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] "Suicide rates, Global Health Observatory (GHO) data," 2015, http://www.who.int/gho/mental_health/suicide_rates/en/.

[2] V. Venek, S. Scherer, L. P. Morency, A. S. Rizzo, and J. Pestian, "Adolescent suicidal risk assessment in clinician-patient interaction," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 204–215, 2017.

[3] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT Press, 2012.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[7] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 785–794, San Francisco, CA, USA, 2016, ACM.

[8] S. Chattopadhyay, "A mathematical model of suicidal-intent-estimation in adults," *American Journal of Biomedical Engineering*, vol. 2, no. 6, pp. 251–262, 2012.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] R. C. O'Connor and M. K. Nock, "The psychology of suicidal behaviour," *The Lancet Psychiatry*, vol. 1, no. 1, pp. 73–85, 2014.

[11] D. Sikander, M. Arvaneh, F. Amico et al., "Predicting risk of suicide using resting state heart rate," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–4, Jeju, Republic Korea, 2016, IEEE.

[12] N. Jiang, Y. Wang, L. Sun, Y. Song, and H. Sun, "An ERP study of implicit emotion processing in depressed suicide attempters," in *2015 7th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 37–40, Huangshan, China, 2015, IEEE.

[13] W. C. Chiang, P. H. Cheng, M. J. Su, H. S. Chen, S. W. Wu, and J. K. Lin, "Socio-health with personal mental health records: suicidal-tendency observation system on Facebook for Taiwanese adolescents and young adults," in *2011 IEEE 13th International Conference on e-Health Networking, Applications and Services*, pp. 46–51, Columbia, MO, USA, 2011, IEEE.

[14] Y. P. Huang, T. Goh, and C. L. Liew, "Hunting suicide notes in web 2.0 - preliminary findings," in *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, pp. 517–521, Beijing, China, 2007, IEEE.

[15] K. D. Varathan and N. Talib, "Suicide detection system based on Twitter," in *2014 Science and Information Conference*, pp. 785–788, London, UK, 2014, IEEE.

[16] J. Jashinsky, S. H. Burton, C. L. Hanson et al., "Tracking suicide risk factors through Twitter in the US," *Crisis*, vol. 35, no. 1, pp. 51–59, 2014.

[17] W. Wang, L. Chen, M. Tan, S. Wang, and A. P. Sheth, "Discovering fine-grained sentiment in suicide notes," *Biomedical Informatics Insights*, vol. 5, Supplement 1, pp. 137–145, 2012.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, https://arxiv.org/abs/1301.3781.

[19] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "DiSAN: directional self-attention network for RNN/CNN-free language understanding," 2017, https://arxiv.org/abs/1709.04696.

[20] S. Chattopadhyay, "A study on suicidal risk analysis," in *2007 9th International Conference on e-Health Networking, Application and Services*, pp. 74–78, Taipei, Taiwan, 2007, IEEE.

[21] Y. M. Tai and H. W. Chiu, "Artificial neural network analysis on suicide and self-harm history of Taiwanese soldiers," in *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*, pp. 363–363, Kumamoto, Japan, 2007, IEEE.

[22] M. Liakata, J. H. Kim, S. Saha, J. Hastings, and D. Rebholz-Schuhmann, "Three hybrid classifiers for the detection of emotions in suicide notes," *Biomedical Informatics Insights*, vol. 5, Supplement 1, 2012.

[23] E. Okhapkina, V. Okhapkin, and O. Kazarin, "Adaptation of information retrieval methods for identifying of destructive informational influence in social networks," in *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pp. 87–92, Taipei, Taiwan, 2017, IEEE.

[24] M. Mulholland and J. Quinn, "Suicidal tendencies: the automatic classification of suicidal and non-suicidal lyricists using NLP," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 680–684, Nagoya, Japan, 2013.

[25] X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li, and T. Zhu, "Detecting suicidal ideation in Chinese microblogs with psychological lexicons," in *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*, pp. 844–849, Bali, Indonesia, 2014, IEEE.

[26] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: a content analysis," *Biomedical Informatics Insights*, vol. 3, pp. 19–28, 2010.

[27] D. Delgado-Gomez, H. Blasco-Fontecilla, F. Sukno, M. Socorro Ramos-Plasencia, and E. Baca-Garcia, "Suicide attempters classification: toward predictive models of suicidal behavior," *Neurocomputing*, vol. 92, pp. 3–8, 2012.

[28] D. Delgado-Gomez, H. Blasco-Fontecilla, A. A. Alegria, T. Legido-Gil, A. Artes-Rodriguez, and E. Baca-Garcia, "Improving the accuracy of suicide attempter classification," *Artificial Intelligence in Medicine*, vol. 52, no. 3, pp. 165–168, 2011.

[29] J. P. Pestian, P. Matykiewicz, M. Linn-Gust et al., "Sentiment analysis of suicide notes: a shared task," *Biomedical Informatics Insights*, vol. 5, Supplement 1, pp. 3–16, 2012.

[30] S. J. Cash, M. Thelwall, S. N. Peck, J. Z. Ferrell, and J. A. Bridge, "Adolescent suicide statements on MySpace," *Cyberpsychology, Behavior, and Social Networking*, vol. 16, no. 3, pp. 166–174, 2013.

[31] A. Shepherd, C. Sanders, M. Doyle, and J. Shaw, "Using social media for support and feedback by mental health service users: thematic analysis of a Twitter conversation," *BMC Psychiatry*, vol. 15, no. 1, p. 29, 2015.

[32] F. Ren, X. Kang, and C. Quan, "Examining accumulated emotional traits in suicide blogs with an emotion topic model," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1384–1396, 2016.

[33] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on Twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.

[34] H. Y. Huang and M. Bashir, "Online community and suicide prevention: investigating the linguistic cues and reply bias," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI'16*, San Jose, CA, USA, 2016.

[35] M. Kumar, M. Dredze, G. Coppersmith, and M. De Choudhury, "Detecting changes in suicide content manifested in social media following celebrity suicides," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, pp. 85–94, Guzelyurt, Northern Cyprus, 2015, ACM.

[36] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering shifts to suicidal ideation from mental health content in social media," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pp. 2098–2110, San Jose, California, USA, 2016, ACM.

[37] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, *The Development and Psychometric Properties of LIWC2015*, University of Texas at Austin, 2015.

[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[39] A. Voutilainen, "Part-of-speech tagging," in *The Oxford Handbook of Computational Linguistics*, pp. 219–232, Oxford University Press, 2003.

[40] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal Component Analysis*, pp. 115–128, Springer, 1986.

[41] S. Pan, J. Wu, and X. Zhu, "CogBoost: boosting for fast cost-sensitive graph classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 2933–2946, 2015.

[42] S. Pan, J. Wu, X. Zhu, G. Long, and C. Zhang, "Task sensitive feature exploration and learning for multitask graph classification," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 744–758, 2017.

[43] S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang, "Adversarially regularized graph autoencoder," 2018, https://arxiv.org/abs/1802.04407.