2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
Banff Center, Banff, Canada, October 5-8, 2017

# Detecting Suicide Ideation from Sina Microblog

Yuanbo Gao*, Baobin Li*, Xuefei Wang†, Jingying Wang†, Yang Zhou†, Shuotian Bai‡ Tingshao Zhu†
*School of Computer and Control University of Chinese Academy of Sciences
Beijing,100190, China Email:libb@ucas.ac.cn
†Institute of Psychology Chinese Academy of Sciences
Beijing, 100101, China Email:tszhu@psych.ac.cn
‡School of Information Engineeringm, Hubei University of Economics Email:baishuotian@hbue.edu.cn

***Abstract*—Suicide is becoming a serious problem, and how to prevent suicide has become a very important research topic. The development of Social Network System (SNS) provides an ideal platform to monitor persons' suicidal ideation. Based on Sina microblog (Weibo), this paper proposes a real-time monitoring system detecting users' suicidal ideation. From 59046 posts collected with labels of either suicide or non-suicide, we extract new features based on content and emotion. Finally, four different classifiers including Support Vector Machine(SVM), Multinomial Naive Bayes(MultiNB), Logistic Regression(LR), Multi Layer Perception(MLP) are used to construct classified model respectively. Experimental results show that it is possible to detect suicidal ideation from microblogs, and the result of MLP is the best, whose F1 is up to 67.6%.**

***Index Terms*—Suicide-ideation, Social Network, Sina microblog, LSVM, LR, MultiNB, MLP**

## I. Introduction

Suicide is becoming a serious problem, and the suicide rate increases year by year. World Health Organization (WHO) reports that there are 800,000 people kill themselves every year. In China, about two millions people try to suicide, and 5% succeed [1]. Up to 2012, suicide has become the second leading cause of death for young people aged between 15 and 29 [2]. In order to improve public attention to suicide problem, WHO has set September 10 every year as the World Suicide Prevention Day. Suicide has become the global concern of public health, mental and social problem [3].

It is reported that 90% suicide death are caused by serious psychological problems [4], although there are many other factors that may cause suicide, such as physical, social, or cultural factors, etc. For patients with mental disorders, such as depression and schizophrenia, are most likely to have a tendency to commit suicide [5]. Usually, people who are considering suicide are suffering from a combination of poor mental health and difficult life event. It has been estimated 50% mental illness patients tried to commit suicide [6], [7]. If we can detect suicidal ideation in time, we may provide help effectively to reduce the occurrence of suicide.

The traditional method to identify suicidal ideation is self-report by using questionnaire, such as Questionnaire of Suicide Attitude (QSA), Beck Suicide ideation Scale (BSS) [8], etc. While questionnaire has been widely used with high validity, it still has several disadvantages. First, self-report cannot acquire real-time changes of person's ideation, and people may get bored to answer too many questions frequently. Second, the sample size by using questionnaire is limited, and it can not be conducted in large scale. In addition, people may not report his/her real feeling and give inauthentic answers deliberately. This would decrease the validity of the results. In this paper, we propose to identify suicidal ideation from microblogs published on Sina. Moreover, due to rapid development of computer science, data mining and text analysis, it becomes possible that monitoring suicidal ideation automatically.

Recently, with the help of social medias and machine learning, researchers can collect large amount of data and automatic detect users who have suicidal ideation to overcome traditional problems as soon as possible. There have many researches based on social medias both in China and board. Foreign researches are mainly based on Twitter and Facebook [9]–[12] while in China, Sina microblog has become a popular communicating platform where people can get quantitative data, which is a basis of recognizing suicidal ideation, and a lot of relative works have been finished [13]–[15].

For previous researches, they supposed that the suicidal ideation microblogs are posted by some people which have been confirmed death, but after checking carefully, we find out some microblogs have no suicidal thought, and the number of these suicidal ideation data which have been used in previous studies is small, only 614 [13] and 664 [14]. Second, they only focused on features of formal language (e.g. regular sentimental words), rather than features of informal language (e.g. cyberwords). On social medias, users prefer to expressing themselves freely, thus, we need to examine characteristics of informal language among suicidal users. Second, previous studies rarely examine the semantic relationships between words. Ignoring the semantic relationship may lead to wrong predicting results.

In this paper, we aim to establish a suicidal ideation predicted model based on Sina Microblog, and the process is shown in Fig 1. First, Sina Microblog are obtained by Crawler though Open-Platform API [16]. Second, traditional word segmentation(stanford NLP [17])is applied on each data. Third, we remove special stopwords of Sina Microblog which has been segmented and use N-Gram to combine adjacent tokens in order to capture the language structure from the statistical point of view. Then, we extract content and emotion features, and construct model. Finally, ten-fold cross validation is used to evaluate the model and the outstanding performance prove that our method is effective. A brief introduction of our contribution is below:

1)New microblogs are labeled by professional and work out a standard, 9123 suicide and 49923 non-suicide data, which ensure the data is enough big and correct. 2) For text analysis, we consider cyberwords for the first time. Through analysing data, we find that people who have suicidal ideation use cyberwords less frequently. 3) Synonyms dictionary is designed in order to consider the semantic connection between words. 4) We collect the suicide emotion dictionary and adopt a new segmentation method for Sina Microblog.

## II. Related Work

At present, social networks are becoming popular all over the world, such as Twitter and Facebook. Every day, there are huge of comments and tweets published, and people prefer to expressing their feelings online. Jo Robinson et al. [9] found that social media behavior was related to real behavior. It is feasible that using social media data to monitor the public suicidal ideation [10]. We can take advantage of the association between suicide-related tweets and suicidal behavior, to identify suicidal young people on the Internet [11]. O'Dea *et al.* [12] analyzed the level of concern among suicide-related tweets, using both human coders and an automatic machine classifier.

Sina Microblog, a popular Chinese social network platform, also provides valuable data for identifying suicidal ideation. By the end of 2014, there has been 53 suicide cases exposed in Sina Microblog. The data of people who have committed suicide could be collected as experimental data with label of suicide. Based on these data, many researches have been conducted. Huang *et al.* (2014) [13] established suicidal prediction models using several classical classifiers, such as Naive Bayes, Logistic Regression, J48, Random Forest, SMO, SVM from LibSVM [18]. Ten-fold cross-validation was utilized to evaluate these models and SVM classifier got the best performance with F1 68.3%, Precision 78.9%, and Recall 60.3%. Huang *et al.* (2015) [14] improved his work by introducing the Topic Model. The modified model got the best result on topic-500 with F1 80.0%, Precision 87.1%, Recall 73.9%, and Accuracy 93.2%. Ang Li *et al.* [15] analyzed the Sina Microblog of 1785 users who expressed their suicidal ideation directly and publicly, and built a subjective well-being prediction model. However, because of taking the traditional psychology scales as golden standard, the methods mentioned may be influenced by sampling bias.

## III. Data Collection and Feature Extraction

By calling Open-Platform API, we download microblogs by using Python-based crawler [16]. We find many comments under one special Microblog user, expressing depression emotion, despair thought and suicidal ideation. So we just download all of these comments from March to September, 2016. A total of 59999 pieces of microblogs are downloaded, and five graduate students specialised in suicide prevention from Institute of psychology, Chinese Academy of Sciences, code these microblogs. There are three rules to code: 1) Expressing the idea of suicide or despair but have no specific action; 2) Having suicide plan but do not need to intervene immediately, which just talked about the way, tool, time and place of suicide or the author have tried to suicide or self-injury in the past, and now the author still have suicidal thought; 3) In great risk and need to intervene immediately.

In the first instance, five researchers classify a small random set of microblogs (n = 500), $\chi^2$ tests with an alpha level of 0.05 are used to compare differents. Kendall's W [19] coefficient is calculated to measure the level of agreement among these researchers, but the agreement($\chi^2 = 1746.08, k = 0.7$) of first round is not good enough. After discussing carefully, five researchers choose randomly another 500 microblogs to code again, the agreement ($\chi^2 = 2112.57, k = 0.85$)is improved and satisfactory. Then, the rest microblogs are divided into five parts to five researchers for coding. In the coding duration, an additional two options are created: 'data known' (the microblog can be understand) and 'data discard' (the microblog cannot be understood; used sparingly and does not include cases where the context is simply ambiguous). As such, the five researchers complete the final coding task on a large sample of microblogs (n = 59046), including 9123 microblogs with suicidal ideation, and 49923 without suicidal ideation.

### *A. Data Preprocessing*

We use Stanford CoreNLP Parser [17] to segment each microblog. But after carefully investigating the results, we find many phrases are parsed incorrectly. So N-gram is used to combine adjacent words to reduce the error rate of segmenting. N-gram is contiguous sequence of n-item from a given sequence of text, the final result of preprocessing for every microblog is composed of uni-gram, bi-gram and tri-gram.

### *B. Dictionary Construction*

We run text analysis to extract linguistic features from text [20]. In this paper, we develop some dictionaries including Stop words, Suicide lexicon, Cyberspeak and Synonymous dictionary for features extraction.

*1) Stopwords dictionary:* Based on our datasets collected, we create a stop words dictionary to filter out the punctuation and common words. The commonly used word is the one often appeared in microblogs, but be worthless for suicidal ideation prediction, such as "I, forward".

*2) Suicide lexicon dictionary:* There are varieties of phrase to express suicide ideation, in particular for Sina Microblog. It is necessary to collect special suicide lexicon words which can be regraded as an important feature. So, we construct the lexicon dictionary based on existing suicide Microblog comments, and select suicide phrases to construct suicide lexicon dictionary such as "gas, death, charcoal-burning".

*3) Cyberspeak dictionary:* With the popularity of the Internet, there are more and more cyberspeak. The big difference between cyberspeak and other words is that cyberspeak contain more strong emotions. Thus, the cyberspeak often be used for banter, ridicule and entertainment instead of being used in very serious situation. Usually, a suicidal microblog is

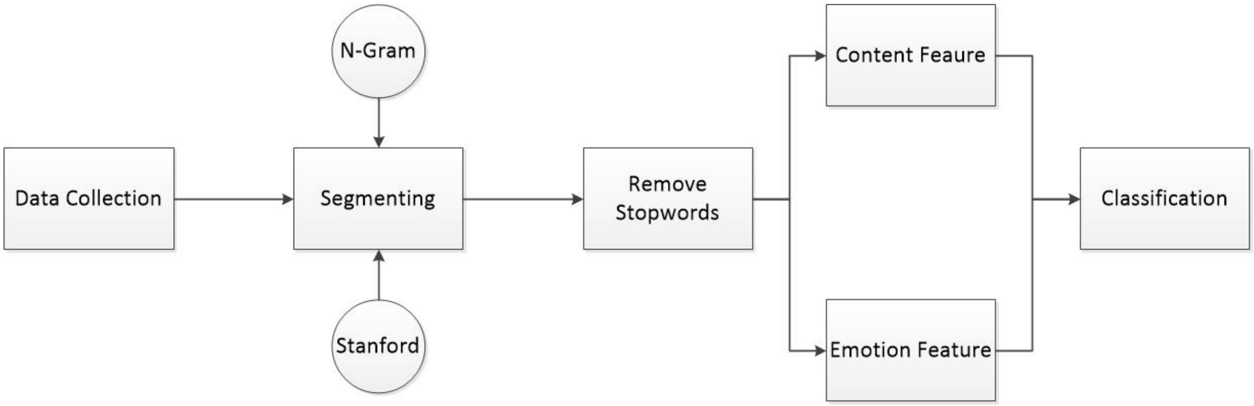

Fig. 1. The process of suicide ideation prediction

published by a person who may be in extreme melancholy situation and it is impossible for him to use the cyberspeak ordinarily. So, the frequency of using cyberspeak could be considered as one of classification attributes. Currently, there is no a cyberspeak dictionary, so based on the existing social network data, we collect these commonly used cyberspeak and construct a Cyberspeak dictionary.

*4) Synonymous dictionary:* In order to reduce the redundant information and consider the semantic similarity between words in these microblogs, we built a synonymous dictionary. According this dictionary, different words of the same or similar meaning would all be replaced by the same word. For example, the meanings of “happy”, “glad”, “delighted” and “excited” are similar. Based on our synonymous dictionary, in the text, they would all be replaced by “happy”.

In the existing synonymous dictionary, the synonyms are recorded in pairs. For example, A:B means the word A and word B are a pair of synonyms, and this combination just be recorded once in dictionary. In addition, the word A or word B may be recorded in the dictionary with another synonymous words in pair. There are 42536 pairs synonyms in the dictionary in total.

Considering one word may be recorded many times, it is not convenient for calculation, so we try to merge these relevant synonymous words together. In the merger process, in order to avoid expanding or missing the synonymity of words, we conduct two transfers processing for the original dictionary. First, we merge lines which have the same key word in synonymous dictionary. Second, based on the result of first step, we find the synonymous words between lines. Finally, we get a new synonymous dictionary with 7505 sets synonyms. The comparison of original version and modified version is shown in Table I.

### C. Features Extraction

In this section, we mainly extract two types of features: content and emotion features. Content features focus on the term-frequency and inverse document frequency of key words, and emotion features are the ratios of cyberwords and suicide words.

TABLE I
THE COMPARISON OF ORIGINAL VERSION AND MODIFIED VERSION

| | |
|---|---|
| Original synonyms pairs | 42536 |
| Current synonyms set | 7505 |

We utilize Term Frequency-Inverse Document Frequency(TF-IDF) [21] to extract content features. TF-IDF is a kind of statistic method which is used to evaluated the important degree of a word to a set of files or a corpus. In this study, concretely, four steps are conducted as following:

1) We find out keywords according to our database, the probability of uni-gram (1), bi-gram (2), tri-gram (3) are computed among the suicide data respectively.

$$p(x_i) = \frac{count(x_i)}{\sum_i count(x_i)} \tag{1}$$

$$p(x_{i+1}|x_i) = \frac{count(x_i, x_{i+1})}{count(x_i)} \tag{2}$$

$$p(x_{i+2}|x_i, x_{i+1}) = \frac{count(x_i, x_{i+1}, x_{i+2})}{count(x_i, x_{i+1})} \tag{3}$$

Then, we sort these words in accordance with the probability and made a synonym substitution(the low probability of words to be replaced). The top 2000 words are chosen as keywords.

2)For each microblog, we calculate the value of term frequency of each keyword. If the keyword is not in this microblog, the value of term frequency inverse document frequency of the keyword is 0 and skip the following step. We describe the process of computing term frequency at first. The number of each keyword’s occurrence is calculated in this microblog, and we normalize these frequencies to make data more smooth as the equation (4). $tf_{ij}$ denotes the value of term frequency of word, where $i$ and $j$ mean the $i$th keyword in $j$th mircoblog, and $\overline{tf_{ij}}$ means the result of data normalization for the microblog.

$$\overline{tf_{ij}} = \frac{tf_{ij}}{max(tf_{ij})} \tag{4}$$

3)Second, the value of inverse document frequency of each keyword is calculated as the equation (5). In which, $N_b$ denotes the total number of microblogs, and $N_{t_i \in d_j}$ denotes the number of all of microblogs which contain the keyword.

$$idf_i = \log \frac{N_b}{N_{t_i \in d_j}} \tag{5}$$

4)At last, the weight of each keyword is calculated based on the equation (6), which act as content features for this microblog.

$$w_{ij} = \overline{tf_{ij}} \times idf_i \tag{6}$$

In addition, we calculate the ratios of suicide emotion words and cyberwords separately as emotion features for this microblog.

## IV. Results

Based on above content and emotion features. We use cross-validation method, the average accuracy when the training set is divided into 10 ”folds” or subsets is assessed, with each fold being used as an intermediate testing subset for the other 9 folds. The performance of classification are measured by Accuracy as well as Precision, Recall, F1 metrics. Precision refers to the ratio of true suicidal microblogs against all microblogs predicted as suicidal. Recall refers to the fraction of suicidal instances retrieved by trained models. Accuracy refers to all predictions match their labels regardless whether they are suicidal microblogs or not. F1 is the harmonic mean of the two and represents a balance between the two. The range of the precision, recall and the F1 metrics are all bounded between 0 and 1, of which a higher value indicates better performance. These metrics are defined as:

$$Precision = \frac{correctly\ identified\ items\ of\ suicidal}{suggested\ items\ of\ suicidal} \tag{7}$$

$$Recall = \frac{correctly\ identified\ items\ of\ suicidal}{actual\ items\ of\ suicidal} \tag{8}$$

$$F1 = \frac{2 \times Recall \times Precision}{(Recall + Precision)}. \tag{9}$$

The suicidal ideation predicted models are constructed with SVM(linear kernel), LR, MultiNB, and MLP. In this paper, we firstly try MultiNB and MLP to detect suicidal ideation.

1) Multinomial Naive Bayes: Naive Bayes is a conditional probability model, it makes the assumption that the value of a particular feature is independent of the value of any other feature, given the class variable. The common decision rule of Naive Bayes is known as maximum a posteriori. We use multinomial naive bayes classifier, which is suitable for discrete features.

2) Multi-Layer Perception: we use two hidden layers, each of which contain 1000 neurons, this choice is based on previous work in which it has been shown that the use of two or more hidden layers has a marginal effect on the network performance [22]. The output layer is designed to follow the standard encoding of labels. Stochastic gradient descent is applied to update parameters and the initial learning rate is 0.1.

Table II presents corresponding results of precision, recall, F1 for our model. Through comparison, The performance of MLP is relative better. The Accuracy(92.3%) and F1 (67%) of MLP are the highest. The precision of MultiNB is up to 70% and the recall of SVM is 77%. The outstanding results show that our model would be a great help for suicide prevention, which could help psychologists to identify suicide posts and understand of suicides' inner activities effectively.

We have implemented this prediction model into our suicide monitoring system, which is still in its early stage. The aim of the suicide monitoring system is to help professionals conduct suicide intervention in time. It can obtain real-time data from Sina microblog, and predict each user's suicide ideation based on his/her microblogs by our model. The framework of suicide monitoring system is shown in Fig. 2.

## V. Discussion

This paper mainly focus on whether the suicide ideation expressed in microblogs can be identified by machine-learned classifier. The machine learned classifier correctly identified 77% of suicide-ideation sina microblogs and achieve an overall agreement rate of 92.3%.

We use N-gram to extract tf-idf features and emotion features, which have been already used in previous studies [12]–[14]. In this paper, given the suicide ideation might be expressed on the web differing from in real-life, we construct suicide-related emotion dictionary based on Sina Microblog. Moreover, we find the suicide-related microblogs are normal in general, where the probability of using cyberspeak is small, so the frequency of cyberspeak can act as a important cue to recognize suicide-related microblogs. The relationship between words are also considered in our study, and similar words are replaced with high frequency word, which can reduce sparsity and eliminate ambiguity to some extent. The classification results show that these dictionaries we build are valid and 54499 microblogs are classified correctly. These findings illustrate a significant advancement in our ability to reliably detect suicide risk in social media data.

A real-time monitoring system is designed by using model which is constructed by machine-learned classifier. This system can feedback suicide-ideation of Microblog users in time. We will send private messages to help these users who have suicide-ideation, but the ethics of this type of suicide detection remain difficult to navigateusers must consent to their microblogs being monitored by an organisation or an individual, and permission to be contacted if a suicide-ideation microblog is detected. Future research must consultation with

TABLE II
THE COMPARISON RESULTS OF DIFFERENT FEATURES

| Features | Classifier | Accuracy | F-measure | Precision | Recall |
|---|---|---|---|---|---|
| Content | MultiNB | 86.3% | 61.3% | 54.5% | 70.1% |
| | LR | 88.9% | 57.7% | 70.7% | 48.7% |
| | SVM | 88.8% | 56.8% | 71.1% | 47.3% |
| | MLP | 90.5% | 64.7% | 71.2% | 59.3% |
| Content+Emotion | MultiNB | 86.4% | 61.7% | 54.9% | **70.4%** |
| | LR | 89.9% | 62.1% | 74.1% | 53.4% |
| | SVM | 90.7% | 65.4% | **77%** | 56.9% |
| | MLP | **92.3%** | **67.6%** | 76.2% | 60.8% |

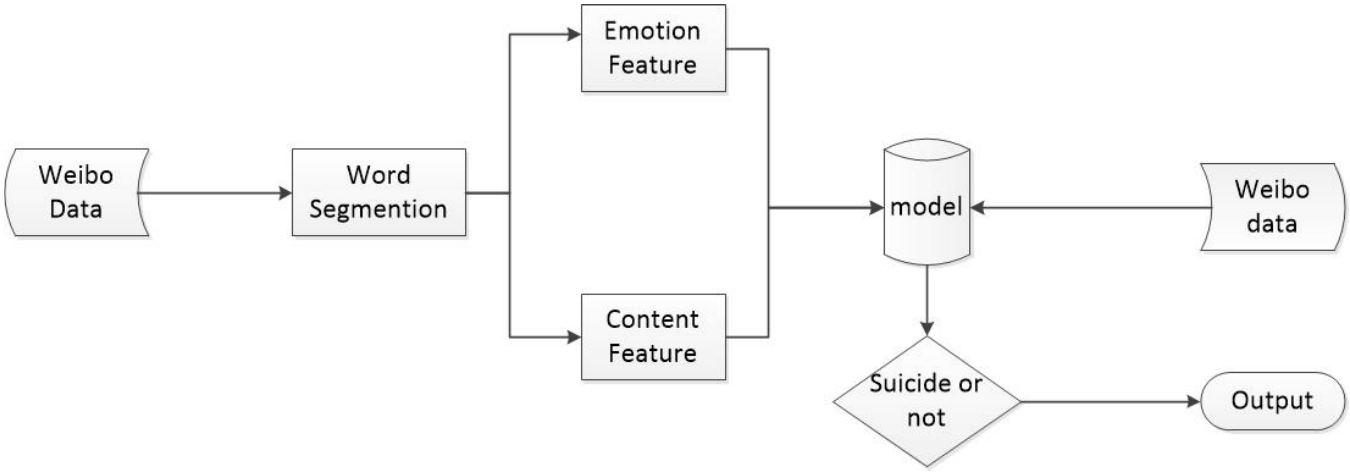

Fig. 2. The framework of suicide monitoring system

Sina Microblog users, consumers and mental health professionals before such a tool could be considered for public use.

In addition, our work also exist some limitations.

- It cannot be definitively stated that all suicide-related microblogs are genuine statements of suicidality or that the suicide-related microblogs collected are truly indicative of suicide. Nonetheless, these statements evoked of a strong level of concern among coders and are believed to warrant further investigation.
- It is difficult to obtain the context beyond the users' post automatically and without direct contact with the account holder. Future track and analysis of the replies to the some special users' posts may help to clarify the level of risk.
- The accuracy of suicidal posts is up to 77% in this paper, However, 23% of suicidal posts are not identified correctly . This may mainly due to the complexity and ambiguity of Chinese sentences on the Internet and the unbalance between suicide and non-suicide posts.

## VI. CONCLUSION

Suicide is not just a personal issue, but also a social event that deserves public attention and intervention. Identification of suicidal ideation is just the first step, which is also the most important step for suicide intervention. Predicting user's suicidal ideation based on social media data could avoid the problems of non-real-time and subjectivity caused by the traditional self-report method. In this paper, we propose to identify suicidal ideation based on microblogs, and we construct our own dictionary based on Sina Microblog. The effective features are extracted from microblogs including content and emotion features. Ten-fold cross validation are used to evaluate the performance of trained model. The experimental results indicate that the highest accuracy of suicidal ideation prediction is 92.3%. Moreover, we develop a real-time suicide monitoring system, to analyse hundreds of millions users' microblogs, and assess Microblog users' suicidal ideation in time.

Future works would extract more features and apply more advantages text analysis methods to further mining latent social relationships, which can overcome the vague items in microblogs as soon as possible. Topic Model [23] is a very useful tool to analyse text. It may find latent semantics in suicide posts, and it also can find dependency on suicide emotion dictionary. We plan to use latent social network relationships in suicide individuals and groups. People who have interactions in their posts might propagate information to others. Some research have been done on this topic [24], and shed lights on this part in media propagation. In addition, in the sentence embedding usually the relationship among words in the sentence, i.e., the context information, is taken into consideration, and the time-delay networks(long short-term memory(LSTM)) would also used to build better model, which have been widely used in text analysis and classification [25], [26]. The range of suicide-related search terms would be expanded, through coding more suicide-ideation posts, to balance microblogs and augment dictionaries.

## VII. Acknowledgments

This work was supported in part by NSFC under Grant 11301504,in part by National Basic Research Program of China(2014CB744600), and in part by CAS Strategic Priority Research Program (XDA06030800) and Natural Science Foundation of Hubei Province (2016CFB208) .

## VIII. References

### References

[1] S. Goldsmith, T. Pellmar, A. Kleinman, and W. Bunney, "Institute of medicine (us), committee on pathophysiology and prevention of adolescent and adult suicide," *Reducing suicide: a national imperative*, 2002.

[2] W. H. Organization *et al.*, *Preventing suicide: A global imperative*. World Health Organization, 2014.

[3] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle, "Tracking suicide risk factors through twitter in the us," *Crisis*, 2015.

[4] B. Barraclough, J. Bunch, B. Nelson, and P. Sainsbury, "A hundred cases of suicide: clinical aspects," *The British Journal of Psychiatry*, vol. 125, no. 587, pp. 355–373, 1974.

[5] D. S. Vandivort and B. Z. Locke, "Suicide ideation: its relation to depression, suicide and suicide attempt," *Suicide and life-threatening Behavior*, vol. 9, no. 4, pp. 205–218, 1979.

[6] T. E. J. Jr., J. S. Brown, and L. R. Wingate, "The psychology and neurobiology of suicidal behavior," *Psychology*, vol. 56, no. 56, pp. 287–314, 2005.

[7] A. Mcgirr, J. Renaud, M. Seguin, M. Alda, C. Benkelfat, A. Lesage, and G. Turecki, "An examination of dsm-iv depressive symptoms and risk for suicide completion in major depressive disorder: a psychological autopsy study." *Journal of Affective Disorders*, vol. 97, no. 1-3, pp. 203–9, 2007.

[8] D. J. Dozois and R. Covin, "The beck depression inventory-ii (bdi-ii), beck hopelessness scale (bhs), and beck scale for suicide ideation (bss)." 2004.

[9] J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, and H. Herrman, "Social media and suicide prevention: a systematic review," *Early Interv Psychiatry*, 2015.

[10] S. Bai, R. Gao, and T. Zhu, "Determining personality traits from renren status usage behavior," in *Computational Visual Media*. Springer, 2012, pp. 226–233.

[11] H. Sueki, "The association of suicide-related twitter use with suicidal behaviour: a cross-sectional study of young internet users in japan," *Journal of affective disorders*, vol. 170, pp. 155–160, 2015.

[12] 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.

[13] X. Huang, L. Zhang, T. Liu, D. Chiu, T. Zhu, and X. Li, "Detecting suicidal ideation in chinese microblogs with psychological lexicons," *arXiv preprint arXiv:1411.0778*, 2014.

[14] X. Huang, X. Li, L. Zhang, T. Liu, D. Chiu, and T. Zhu, "Topic model for identifying suicidal ideation in chinese microblog," 2015.

[15] L. Ang, H. BiBo, B. ShuoTian, and Z. TingShao, "Predicting psychological features based on web behavioral data:mental health status and subjective well-being," *SCIENCE CHINA PRESS*, vol. 11, p. 006, 2015.

[16] Sina, "Sina weibo open platform," Website, 2014, http://open.weibo.com/.

[17] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit." in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[18] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[19] B. S. Linn, M. W. LINN, and L. Gurel, "Cumulative illness rating scale," *Journal of the American Geriatrics Society*, vol. 16, no. 5, pp. 622–626, 1968.

[20] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[21] G. Chowdhury, *Introduction to modern information retrieval*. Facet publishing, 2010.

[22] M. Holmberg, F. Winquist, I. Lundström, J. Gardner, and E. Hines, "Identification of paper quality using a hybrid electronic nose," *Sensors and Actuators B: Chemical*, vol. 27, no. 1, pp. 246–249, 1995.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[24] Y.-Y. Chen, F. Chen, D. Gunnell, and P. S. Yip, "The impact of media reporting on the emergence of charcoal burning suicide in taiwan," *PloS one*, vol. 8, no. 1, p. e55000, 2013.

[25] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[26] S. Wang and J. Jiang, "Learning natural language inference with lstm," *arXiv preprint arXiv:1512.08849*, 2015.