Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention

Lei Cao[1,2] and **Huijun Zhang**[1,2] and **Ling Feng**[1,2] and **Zihan Wei**[3]
Xin Wang[1,2] and **Ningyun Li**[1,2] and **Xiaohao He**[1,2]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Beijing National Research Center for Information Science and Technology, Beijing, China
[3]School of Software, Beihang University, Beijing, China

`{cao-l17,zhang-hj17,xin-wang18,liny18}@mails.tsinghua.edu.cn,`
`fengling@mail.tsinghua.edu.cn, sy1721114@buaa.edu.cn,`
`hexh17@mails.tsinghua.edu.cn`

Abstract

Despite detection of suicidal ideation on social media has made great progress in recent years, people's implicitly and anti-real contrarily expressed posts still remain as an obstacle, constraining the detectors to acquire higher satisfactory performance. Enlightened by the hidden "tree holes" phenomenon on microblog, where people at suicide risk tend to disclose their inner real feelings and thoughts to the microblog space whose authors have committed suicide, we explore the use of tree holes to enhance microblog-based suicide risk detection from the following two perspectives. (1) We build suicide-oriented word embeddings based on tree hole contents to strength the sensibility of suicide-related lexicons and context based on tree hole contents. (2) A two-layered attention mechanism is deployed to grasp intermittently changing points from individual's open blog streams, revealing one's inner emotional world more or less. Our experimental results show that with suicide-oriented word embeddings and attention, microblog-based suicide risk detection can achieve over 91% accuracy. A large-scale well-labelled suicide data set is also reported in the paper.

1 Introduction

Suicide is a growing problem in today's society. Each year nearly 800,000 people worldwide commit suicide, which is one person every 40 seconds, and there are many more who attempt it (Organization et al., 2014). Suicide prevention will conduce to human's well-being, of which timely sensing suicide ideation is an essential task.

Existing Solutions. Traditional suicide risk assessment like Suicide Probability Scale (Bagge and Osman, 1998), Adult Suicide Ideation Questionnaire (Fu et al., 2007), Suicidal Affect-Behavior-Cognition Scale (Harris et al., 2015), etc. requires respondents to either fill in a ques-
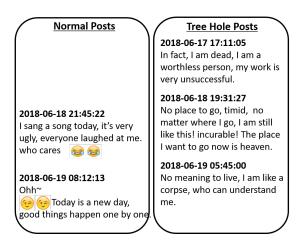

Figure 1: An example of one user's normal posts vs. his/her hidden tree hole posts on microblog.

tionnaire or participate in a professional interview. However, they are applicable to a small group of people. Particularly for the people who are suffering but tend to hide inmost thoughts and refuse to seek helps from others, the approaches cannot function (Essau, 2005; Rickwood et al., 2007).

Recently, the penetration of social media (like forums and microblogs) and its large-scale, low-cost, and open advantages enable researchers to overcome the limitation and timely detect individual's suicide ideation. Despite great efforts have been made (Alambo et al., 2019; Cheng et al., 2017; Du et al., 2018; Sawhney et al., 2018; Coppersmith et al., 2018; Viouls et al., 2018), the social media based detection performance is constrained due to implicitly and anti-real contrarily expressed posts from people who hide their inmost feelings and thoughts on social media. To illustrate, let's see a user's normal blogs vs. his/her hidden posts in a microblog tree hole in Figure 1. Usually, people of suicidal tendency (referred to as suicidal people in the study) tend to disclose their real inner feelings on the microblog space whose authors have committed suicide. Hundreds of such

	Normal Posts	Tree Hole Posts
Avg proportion of self-concern words per post	14%	50%
Avg proportion of others-concern words per post	68%	12%
Avg proportion of suicide-related words per post	5%	95%
Avg number of posts per user in a year	69.3	52.1
Total number of posts from all users in a year	252,901	190,087

Table 1: Statistics of suicidal users' normal posts and hidden tree hole posts on Microblog based on 3652 users from May 1, 2018 to April 30, 2019.

tree holes exist on Sina microblog. An example tree hole contains 1,700,000 posts from suicide attempts. In Figure 1, we can sense a severe hopelessness from the tree hole posts, but not from the normal posts, and the user even expresses an uplift feeling. After cross-examining suicidal users' normal and hidden posts as shown in table 1, we discover that users' hidden posts in the tree hole have more self-concerns, less others-concerns, and illustrate suicidal thoughts more directly. In comparison, suicidal users normal posts contain much less suicide-related words, and the users are reluctant to show their suicidal feelings in their normal posts. Moreover, the data of self-concern and others-concern shown in the first two rows even indicate that people with suicide risk are not willing to talk about themselves in their normal posts, which takes great challenges to detect suicide risk from users' open normal posts.

Our Work. The aim of the study is to break through the above limitation to achieve a new state-of-art performance on latent suicide risk detection from one's open normal microblogs. We leverage tree hole posts from the following two perspectives. (1) We construct suicide-oriented word embeddings based on tree hole contents to strength the sensibility of suicide-related lexicons and context based on tree hole contents. (2) A two-layered attention mechanism is deployed to grasp intermittently changing points from individual's open blog streams, revealing one's inner emotional world more or less. Our experimental results on 252,901 open normal microblogs show that with suicide-oriented word embeddings and two-layered attention, latent suicide risk detection can achieve over 91% accuracy.

In summary, the paper makes the following contributions.

- We build effective suicide-oriented word embeddings to better understand the implicit meanings of words contained in users' nor-

mal posts, and propose a two-layered attention model to capture the changing points which reveal suicide risk from individuals' blog streams. Our latent suicide risk detection from users normal posts not only outperforms the state-of-the-art approaches, but also are powerful enough in detecting implicitly and anti-real contrarily expressed posts.

- We construct a large-scale data set from 3652 suicidal people in the period of [May 1, 2018 to April 30, 2019], containing 252,901 normal posts on Sina microblog. The data set can further facilitate people's well-being studies in the computer science and psychology fields.

2 Related Work

2.1 Traditional Questionnaire-based Suicide Risk Assessment

Researchers have developed a number of psychological measurements to access individual's suicide risk (Pestian et al., 2017), such as Suicide Probability Scale (SPS) (Bagge and Osman, 1998), Depression Anxiety Stress Scales-21 (DASS-21) (Crawford and Henry, 2003; Henry and Crawford, 2005), Adult Suicide Ideation Questionnaire (Fu et al., 2007), Suicidal Affect-Behavior-Cognition Scale (Harris et al., 2015), functional Magnetic Resonance Imaging (fMRI) signatures (Just et al., 2017), and so on. While these measurements are professional and effective, they require respondents to either fill in a questionnaire or participate in a professional interview, constraining its touching to suicidal people who have low motivations to seek help from professionals (Essau, 2005; Rickwood et al., 2007; Zachrisson et al., 2006). A recent study found out that taking a suicide assessment may bring negative effect to individuals with depressive symptoms (Harris and Goh, 2017).

2.2 Suicide Risk Detection from Social Media

Recently, detection of suicide risk from social media is making great progress due to the advantages of reaching massive population, low-cost, and real-time (Braithwaite et al., 2016). Harris et al. (2014) reported that suicidal users tend to spend more time online, have greater likelihood of developing online personal relationships, and greater use of online forums.

Suicide Risk Detection from Suicide Notes. Pestian et al. (2010) built a suicide note classifier used machine learning techniques, which performs better than human psychologists in distinguishing fake online suicide notes from real ones. Huang et al. (2007) hunted suicide notes based on lexicon-based keyword matching on MySpace.com (a popular site for adolescents and young adults, particularly sexual minority adolescents with over 1 billion registered users worldwide) to check whether users have an intent to commit suicide.

Suicide Risk Detection from Community Forums. Li et al. (2013) applied textual sentiment analysis and summarization techniques to users' posts and posts' comments in a Chinese web forum in order to identify suicide expressions. Masuda et al. (2013) examined online forums in Japan, and discovered that the number of communities which a user belongs to, the intransitivity, and the fraction of suicidal neighbors in the social network contributed the most to suicide ideation. De Choudhury et al. (2016) built a logistic regression framework to analyze Reddit users' shift tendency from mental health sub-communities to a suicide support sub-community. heightened self-attentional focus, poor linguistic coherence and coordination with the community, reduced social engagement and manifestation of hopelessness, anxiety, impulsiveness and loneliness in shared contents are distinct markers characterizing these shifts. Based on the suicide lexicons detailing suicide indicator, suicide ideation, suicide behavior, and suicide attempt, Alambo et al. (2019) built four corresponding semantic clusters to group semantically similar posts on Reddit and questions in a questionnaire together, and used the clusters to assess the aggregate suicide risk severity of a Reddit post.

Suicide Risk Detection from Microblogs. Jashinsky et al. (2014) used search keywords and phrases relevant to suicide risk factors to filter potential suicide-related tweets, and observed a strong correlation between Twitter-derived suicide data and real suicide data, showing that Twitter can be viewed as a viable tool for real-time monitoring of suicide risk factors on a large scale. The correlation study between suicide-related tweets and suicidal behaviors was also conducted based on a cross-sectional survey (Sueki, 2015), where participants answered a self-administered online questionnaire, containing questions about Twitter use, suicidal behaviour, depression and anxiety, and demographic characteristics. The survey result showed that Twitter logs could help identify suicidal young Internet users.

Based on eight basic emotion categories (joy, love, expectation, anxiety, sorrow, anger, hate, and surprise), Ren et al. (2015) examined three accumulated emotional traits (i.e., emotion accumulation, emotion covariance, and emotion transition) as the special statistics of emotions expressions in blog streams for suicide risk detection. A linear regression algorithm based on the three accumulated emotional traits was employed to examine the relationship between emotional traits and suicide risk. The experimental result showed that by combining all of three emotion traits together, the proposed model could generate more discriminative suicidal prediction performance.

Natural language processing and machine learning techniques, such as Latent Dirichlet Allocation (LDA), Logistic Regression, Random Forest, Support Vector Machine, Naive Bayes, Decision Tree, etc., were applied to identify users' suicidal ideation based on their linguistic contents and online behaviors on Sina Weibo (Guan et al., 2014; Zhang et al., 2014a; Huang et al., 2014; Zhang et al., 2014b; Huang et al., 2015; Guan et al., 2015; Cheng et al., 2017) and Twitter (Abboute et al., 2014; Burnap et al., 2015; O'Dea et al., 2015; Coppersmith et al., 2015). Deep learning based architectures like Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory Neural Network (LSTM), etc., were also exploited to detect users' suicide risk on social media (Du et al., 2018; Sawhney et al., 2018; Coppersmith et al., 2018). Viouls et al. (2018) detected users' change points in emotional well-being on Twitter through a martingale framework, which is widely used for change detection in data streams.

Category	Suicide Ideation	Suicide behavior	Psychache	Mental illness	Hopeless
Number	586	88	403	48	188
Words/phrases	want to die escape	seppuku hypnotics	want to cry loneliness	depression hallucination	dead end despair

Table 2: Representative words/phrases of Chinese suicide dictionary

3 Suicide-oriented Word Embeddings

Although there are some good works on word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Joulin et al., 2016; Devlin et al., 2018), lack of domain information limits their performance on suicide detection.

Given a serious of pre-trained word embeddings and suicide-related dictionary, we aim to generate suicide-related word embeddings which can strengthen the sensibility of suicide-related lexicons and context. In this study, we call them Suicide-oriented Word Embeddings, as we take advantage of the information from Tree Hole's data set which can be regarded as a kind of latent emotional expressions of individuals. As suicidal individuals in social media often use some suicide-related words/phrases in their posts, we employ Chinese suicide dictionary (Lv et al., 2015) to generate suicide domain associated embeddings. The Chinese suicide dictionary analyzes 1.06 million active blog users' posts and lists 2168 words/phrases related to suicidal ideation. These words/phrases belong to 13 categories and each word/phrase is assigned with a risk weight from 1 to 3 which indicates the relevance of suicide. We list 5 representative categories in table 2.

Since pre-trained word embeddings already contain rich semantic information and contextual information, we only need to enrich existing word embeddings with suicide-related information.

We employ a masked classification task to do this. Generally, a sentence should contains suicide-related words/phrases if it express suicidal ideation. Hence, We select 10,000 sentences[1] from Tree Hole's data set and ensure every sentence contains more than one word/phrase appeared in Chinese suicide dictionary.

Moreover, we utilize the selected sentences to do a suicidal expression classification. A sentence is regarded as suicidal expression only if it in-
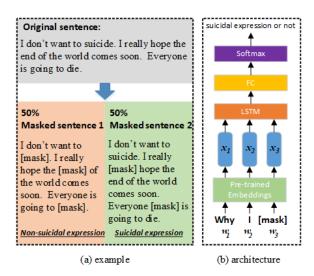

(a) example   (b) architecture

Figure 2: An example of the strategy for embeddings training is shown on the left and the architecture of the masked classification task to train the suicide-oriented word embeddings is on the right.

cludes at least one suicide-related word/phrase. In this way, we do a sentence-level classification to refine pre-trained word embeddings and let them understand which word/phrase is relevant to suicide expression.

In training details, for each epoch, we randomly select 50% sentences to replace all suicide-related words/phrases with "[mask]". Especially, for the rest 50% sentences, we randomly insert two "[mask]" into every sentence to avoid the suicidal expression classifier classifying sentence only based on whether it contains word "[mask]".

An example is given in Figure 2 (a). **Masked sentence 1** is the sentence that we replace all suicide-related words/phrases with "[mask]" and **Masked sentence 2** is the sentence that we randomly insert two "[mask]". We label **Masked sentence 1** as 0 (non-suicide) and **masked sentence 2** as 1 (suicide).

As there is no clear boundary between suicide-related words/phrases and others in pre-trained word embeddings, through this suicidal expression classification we force suicide-related words/phrases to be enriched with domain information and let all the suicide-related word/phrases

---

[1]Through out this work, a "sentence" can be a piece of text, rather than an actual linguistic sentence. It may contains more than one actual sentence.

contain the relationship with suicidal ideation. After classification model converge in Tree Hole's data set, we obtain suicide-oriented word embeddings which contain both semantic information from pre-trained word embeddings and suicide-information from suicide dictionary.

As illustrated in Figure 2, given a sentence $A = \{w_1, w_2, .., w_n\}$ written by a user in Tree Hole, where n is the length of a sentence, the aim of suicidal expression classification is to classify whether this sentence contains expression about suicidal ideation or not. In this case, we define $X = \{x_1, x_2, .., x_n\} \in \mathbb{R}^{n \times d_e}$ as the word embeddings of A, where d_e is the dimension of embeddings. Figure 2 shows the architecture of the suicidal expression classification Model.

We employ a LSTM layer to extract text feature from A followed by a fully connected layer for classification. We feed the word embeddings X into the LSTM as following:

$$h_t = \mathbf{LSTM}(x_i, h_{t-1}),$$
$$[k_1, k_2] = softmax((H^a W_1 + b_1)^T W_2 + b_2) \quad (1)$$

where h_{t-1}, h_t represent the hidden states at time $t-1$ and t, $H^a = \{h_1, h_2, ..., h_n\} \in \mathbb{R}^{n \times d_e}$ is sentence representation of A, $[k_1, k_2]$ stand for the possibility of whether the sentence contains expression about suicidal ideation or not. $W_1 \in \mathbb{R}^{d_e \times 1}$, $W_2 \in \mathbb{R}^{n \times 2}$, $b_1 \in \mathbb{R}^{1 \times 1}$ and $b_2 \in \mathbb{R}^{1 \times 2}$ are trainable parameters.

4 Suicide Detection Model (SDM) based on Suicide-oriented Word Embeddings and Attention Mechanism

Given a sequential of posts $\hat{T}$ from one user, $\hat{T} = \{(s_1, p_1), (s_2, p_2), ..., (s_m, p_m)\}$, where m denotes the number of posts, (s_i, p_i) stand for text and picture from i-th post. The aim is to detect whether the user at risk of suicide or not. Let $\hat{X} = \{x_1, x_2, .., x_n\} \in \mathbb{R}^{n \times d_e}$ be the word embeddings of s_i, where n represents the length of s_i and d_e is the dimension of embeddings. Figure 3 shows the architecture of the proposed two-layered attention model.

4.1 Feature Extraction

Text Feature Extraction. We employ a LSTM layer and attention mechanism to extract text feature from s_i. We feed the word embeddings $\hat{X}$ into
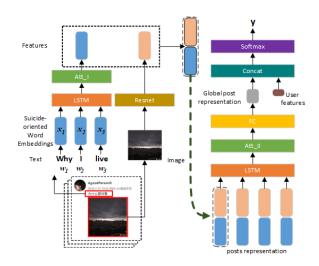

Figure 3: Architecture of the Suicide Risk Detection model.

the LSTM as following:

$$h_t = \mathbf{LSTM}(x_t, h_{t-1}) \quad (2)$$

where h_{t-1}, h_t represent the hidden states at time $t-1$ and t. We obtain the primary textual representation $H_i^p = \{h_1, h_2, ..., h_n\} \in \mathbb{R}^{n \times d_e}$ of s_i after LSTM layer. To gain the critical suicide-related textual information of H_i^p, we apply the **attention mechanism Att_I** :

$$Att\_I = softmax(H_i^p W_3 + b_3) \quad (3)$$

where $Att\_I \in \mathbb{R}^{n \times 1}$ is the attention vector that demonstrates the distribution of the weights for each word of primary textual representation, $W_3 \in \mathbb{R}^{d_e \times 1}$ and $b_3 \in \mathbb{R}^{1 \times 1}$ are trainable parameters. Then we make multiplication between the attention vector Att_I and H_i^p to get the final textual representation $\hat{H}_i \in \mathbb{R}^{d_e \times 1}$ of text s_i,

$$\hat{H}_i = (H_i^p)^T Att\_I \quad (4)$$

Image Feature Extraction. We extract image features from a 34 layer pre-trained ResNet (He et al., 2016). For the convenience of calculation, we convert the last fully connected layer of ResNet from 512×1000 to $512 \times d_e$:

$$I_i = tanh(O W_4 + b_4) \quad (5)$$

where $O \in \mathbb{R}^{1 \times 512}$ is the input of the last fc-layer, $W_4 \in \mathbb{R}^{512 \times d_e}$ and $b_4 \in \mathbb{R}^{1 \times d_e}$ are trainable parameters. Then, $I_i \in \mathbb{R}^{1 \times d_e}$ is the visual representation of picture p_i.

User's Feature Extraction.

Feature	Dimension	Description
Gender	3	(1,0,0)for male, (0,1,0)for female and (0,0,1) for 'Unknown'.
Screen name length	1	The length of screen name.
Post Count	1	The number of posts.
Follower	1	The number of followers.
Following	1	The number of following.
Picture	1	The number of posts with picture.
Post time	4	The four dimensions are proportions of posts posted in (0:00-5:59), (6:00-11:59), (12:00-17:59), (18:00-23:59) of a day.

Table 3: Summary of user's features.

As illustrated in table 3, we extract 12 features $F \in \mathbb{R}^{12 \times 1}$ from user' profile and posting behaviour. Since not every one has tree hole's data, in this study we do not consider tree hole's data in suicide risk detection model.

4.2 Suicide Risk Detection

Given textual representation $\hat{H}_i$ and visual representation I_i, we employ a concatenate operation $\oplus$ to obtain the post representation $E_i \in \mathbb{R}^{2d_e \times 1}$ for a single post (s_i, p_i):

$$E_i = \hat{H}_i \oplus I_i^T \qquad (6)$$

Similar with above, we employ another LSTM layer and attention mechanism to generate global post representation $G \in \mathbb{R}^{30 \times 1}$:

$$\begin{aligned} h_t &= \mathbf{LSTM}(E_i, h_{t-1}), \\ Att\_II &= softmax(H^g W_5 + b_5), \qquad (7) \\ G &= tanh(((Att\_II)^T \times H^g)W_6 + b_6), \end{aligned}$$

where h_{t-1}, h_t represent the hidden states at time $t-1$ and t. $H^g = \{h_1, h_2, ..., h_m\} \in \mathbb{R}^{m \times d_e}$ represents the primary post representation of a user after LSTM layer. As not every post of a user expresses the ideation of suicide, we apply the **attention mechanism Att_II** to gain the high suicide risk post information of H^e. An attention vector $Att\_II \in \mathbb{R}^{m \times 1}$ was computed to present the different risk weight of posts, where $W_5 \in \mathbb{R}^{d_e \times 1}$ and $b_5 \in \mathbb{R}^{1 \times 1}$ stand for trainable parameters. Then based on attention Att_II, we obtain the global post representation G for a user, where $W_6 \in \mathbb{R}^{d_e \times 30}$ and $b_6 \in \mathbb{R}^{1 \times 30}$ stand for trainable parameters.

Finally, we apply a concatenate operation to jointly consider G and F, and through a fully connected layer to compute the possibility of suicide:

	Users	Posts	Posts with image
suicide	3,652	252,901	93,461
non-suicide	3,677	491,130	260,667

Table 4: Statistic of suicide data set.

$$[y_1, y_0] = softmax((G \oplus F)^T W_7 + b_7) \qquad (8)$$

where y_1, y_0 represent the possibility of a user at risk of suicide or not, $W_7 \in \mathbb{R}^{42 \times 2}$ and $b_7 \in \mathbb{R}^{1 \times 2}$ stand for trainable parameters.

5 Experiments

5.1 Data Collection

To make suicide risk detection via social media, we construct two data sets: one from Tree Hole and another from Weibo.

Tree Hole's data set. We studied a suicidal community which exists in the comments of a Chinese student's last posting before the student committed suicide. In March 17, 2012, this student which screen name is "Zoufan" left the last word on Weibo and then committed suicide. For the past seven years, More than 160,000 people gather here and write over 1,700,000 comments which is still continuing to grow. They express their suicidal thoughts, show their tragic experiences and demonstrate their plans of suicide behaviors. In psychology, we can understand this community as a Tree Hole. We crawled all comments from May 1,2018 to April 30, 2019 and selected top 4,000 active users. After that, four doctoral students major in computational mental healthcare were employed to annotate users that whether they are at risk of suicide or not. Specifically, We only decide the user "at suicide risk" based on self-report of

his/her tree hole posts. If a user express clear suicidal thoughts like *"At this moment, I especially want to die. I feel very tired. I really want to be free."* more than 5 times in different days, then we label him/her at suicide risk. Finally we get 190,087 sentences of 3,652 users and the average length per post is 11.96 words.

Suicide data set. To collect users at suicide risk, we crawl user's profile and all created posts in Weibo according to user list from Tree Hole' data set. Besides, we select the users who never submit any post containing expression about suicidal ideation and label them as non-suicide risk. In this case, we discard users whose fans more than 1,500 or posts more than 2,000 because that they may be public figure or organization. The statistic of suicide data set are illustrated in Table 4.

5.2 Data Preprocessing

We carry out the following data preprocessing procedures: 1) **Emoji**. We replace emoji with corresponding word like "happy", "cry" to facilitate our model to understand the emotion of user's post. 2) **URL.** As URL has no use for our detection, we simply remove them from sentences. 3) **Image.** All images posted by users were adjusted to 224×224 for normalized input.

5.3 Experimental Setup

In suicide detection task, we treat recent 100 posts from one user as one sample. After sum up D_2 and D_3, we obtain 7,329 microblog users and training set, validation set and test set contain 6,129, 600, 600 respectively. All sentences are padded to the length of the longest sentence in the data set with word "<PAD>". Batch size is 16 during training process and we use 0.001 as learning rate. Adam Kingma and Ba (2015) is adopted as the optimizer.

We compare suicide-oriented word embeddings with following well-developed word embeddings.

(1) **Word2vec**: The fundamental work for considering the local semantic information of words. We get pre-trained Word2vec word embeddings from Li et al. (2018).

(2) **GloVe**: Context-based unsupervised algorithm, which apply co-occurrence matrix to jointly consider the local and global semantic information. We apply open source tool [2] to train word embeddings on all sentences from Tree Hole's data set.

(3) **Fasttext**: A fast text classification and representation learning model based on Word2vec and hierarchical softmax. Pre-trained FastText word embeddings were obtained from official project [3].

(4) **Bert**: Latest language model based on transformer. We acquire pre-trained Bert model from official project[4] and generate word embedding of each word in a sentence dynamically.

Also, we compare our suicide risk detection model with following well-designed methods.

(1) **LSTM** (Coppersmith et al., 2018): An attention mechanism based Long Short-Term Memory model which can capture contextual information between suicide-related words and others.

(2) **Naive Bayesian (NB) and Support Vector Machine (SVM)** (Pedregosa et al., 2011): Two representative machine learning methods with well-designed features. We use SC-LIWC information (Cheng et al., 2017) as textual features, saturation, brightness, warm/clear color and five-color theme information (Shen et al., 2018) from picture as visual features and user's behaviour features from table 3.

5.4 Results

Three sets of tests were conducted to evaluate the performance of suicide risk detection model with suicide-oriented word embeddings.

5.4.1 Effectiveness of Suicide-oriented Word Embeddings

We compare the performance of LSTM and SDM with seven word embeddings as illustrated in table 5. We find that without suicide-related dictionary, Bert outperforms other three word embeddings with 2% higher accuracy and 1.5% higher F1-score on both models. After leveraging suicide-related dictionary, suicide-oriented word embeddings based on FastText achieves the best performance with accuracy 88.00% 91.33%, F1-score 88.14%, 90.92% on two models. Obviously, there is a gap between suicide-oriented word embeddings and normal word embeddings which can verify the effectiveness of the former.

5.4.2 Effectiveness of Suicide Risk Detection Model

We compare the performance of four model as shown in table 6. In this case, LSTM and SDM

		Word2vec	GloVe	FastText	Bert	So-W2v	So-GloVe	So-FastText
LSTM	Acc(%)	79.21	80.17	82.59	85.15	86.00	86.45	**88.00**
	F1(%)	78.58	79.98	82.18	85.69	86.17	86.69	**88.14**
SDM	Acc(%)	86.54	86.55	87.08	88.89	90.83	91.00	**91.33**
	F1(%)	86.63	85.13	86.91	87.44	90.55	90.56	**90.92**

Table 5: Performance comparison for different word embedding and different detection model, where "So-W2v", "So-Glove" and "So-FastText" represent suicide-oriented word embeddings based on Word2vec, GloVe and Fast-Text respectively. Acc and F1 represent accuracy and F1-score.

(a) *User 1 with* $\rho_{nh} = 0.84$ *and* $\rho_{ah} = 0.88$

(b) *User 2 with* $\rho_{nh} = 0.34$ *and* $\rho_{ah} = 0.91$

(c) *User 3 with* $\rho_{nh} = -0.12$ *and* $\rho_{ah} = 0.6$

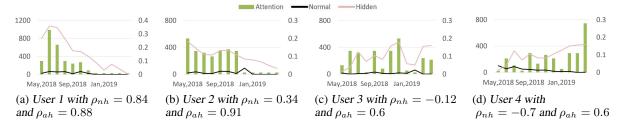(d) *User 4 with* $\rho_{nh} = -0.7$ *and* $\rho_{ah} = 0.6$

Figure 4: Correlation between normal posts, hidden posts and attention on four representative suicide risk users. The x-axis shows dates from May, 2018 to April, 2019. The left y-axis represent number of suicde-related words/phrases and the right represent attention weight.

	Full testset		Harder sub-testset	
	Acc	F1	Acc	F1
SVM	70.34	69.01	61.17	64.11
NB	69.59	70.12	65.14	62.20
LSTM	88.00	88.14	76.89	75.32
SDM	**91.33**	**90.92**	**85.51**	**84.77**

Table 6: Performance comparison between different suicide risk detection model, where Acc and F1 represent accuracy and F1-score respectively.

Inputs	Accuracy	F1-score
Text	88.56	87.99
Text+Image	89.22	89.22
Text+User's feature	90.66	90.17
Text+Image +User's feature	91.33	90.92

Table 7: Ablation test for SDM with different inputs.

employs So-FastText word embeddings as their input. SDM improves the accuracy by over 3.33% and obtains 2.78% higher F1-score on full data set.

5.4.3 Performance on *Harder sub-testset*

To verify the effectiveness of models in dealing with peoples implicit and anti-real contrary expressions on microblog posts, we filter out 130 suicide risk people from test set who do not show obvious suicidal ideation on normal posts. Those 130 people construct a subset of test set named *Harder sub-testset*. After observing the performance of four models as shown in table 6, SDM can keep 8% higher value both in accuracy and F1-score on *Harder sub-testset*, compared with other models, and the decline is smaller than other models. This suggests that SDM performs better than existing models in dealing with people's implicit and anti-real contrary expressions.

5.4.4 Ablation Test for Suicide Risk Detection Model

To show the contribution of different input to the final classification performance, we design an ablation test of SDM after removing different input. All SDMs are based on embedding So-Fasttext. Since not every post contains image and user's features contain missing value, we do not only use images nor user's feature as input of SDMs. As illustrated in table 7, we can see that textual information is a crucial input of our SDM. Besides, user's features play a more important role than visual information. The more modalities we use, the better performance we get.

5.5 Discovery

To further explore the correlation between normal posts and hidden posts from same user, we import *Pearson Correlation Coefficient* (Tutorials, 2014) to manage it. For each user, we obtain a normal vector $V_i^n = \{v_{i,Jan}^n, v_{i,Feb}^n, ..., v_{i,Dec}^n\} \in$

$\mathbb{R}^{12}$, where $v_{i,Jan}^n$ shows the total number of times words/phrases appear in posts for user i in January. In a similar way we get a hidden vector $V_i^h = \{v_{i,Jan}^h, v_{i,Feb}^h, ..., v_{i,Dec}^h\} \in \mathbb{R}^{12}$. For each normal posts, we also have an attention weight from Att_{II} which represents the suicide risk. Then, similar as above, an attention risk vector $V_i^a = \{v_{i,Jan}^a, v_{i,Feb}^a, ..., v_{i,Dec}^a\} \in \mathbb{R}^{12}$ was computed, where $v_{i,Jan}^a$ shows the total suicide risk for user i in January. We donate the pearson correlation coefficient $\rho_{n,h,i}$ of V_i^n and V_i^h, $\rho_{a,h,i}$ of V_i^a and V_i^h as the correlation between normal posts and hidden posts , attention and hidden posts for user i.

As shown in figure 4, we find that there are high positive linear correlation between normal posts and hidden posts from user 1 with $\rho_{nh} = 0.84$ and high negative linear correlation from user 4 with $\rho_{nh} = -0.7$. For other two suicide risk users, there are not obvious linear correlations with $\rho_{nh} = 0.33, -0.12$ respectively. The phenomenon that correlations ρ_{ah} between attention and hidden posts from four users all higher than 0.6 which means high positive linear correlation, which verify the ability of the two-layered attention mechanism to reveal ones' inner emotional world.

6 Conclusion

In this paper, we explore the uses of tree holes to enhance microblog-based suicide risk detection. Suicide-oriented word embeddings based on tree hole contents are built to strengthen the sensibility of suicide-related lexicons and a two-layered attention mechanism is deployed to grasp intermittently changing points from individuals open blog streams. Based on above word embeddings and attention mechanism, we propose a suicide risk detection model which outperforms the well-designed approaches on benchmark data set. Through experimental results we also find that, our model also performs well on people's implicit and anti-real contrary expressions.

Acknowledgments

References

Amayas Abboute, Yasser Boudjeriou, Gilles Entringer, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2014. Mining twitter for suicide prevention. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 250–253. Springer.

Amanuel Alambo, Manas Gaur, Usha Lokala, Ugur Kursuncu, Krishnaprasad Thirunarayan, Amelie Gyrard, Amit Sheth, Randon S Welton, and Jyotishman Pathak. 2019. Question answering for suicide risk assessment using reddit. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 468–473. IEEE.

Courtney Bagge and Augustine Osman. 1998. The suicide probability scale: Norms and factor structure. *Psychological reports*, 83(2):637–638.

Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. 2016. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR mental health*, 3(2):e21.

Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 75–84. ACM.

Qijin Cheng, Tim Mh Li, Chi Leung Kwok, Tingshao Zhu, and Paul Sf Yip. 2017. Assessing suicide risk and emotional distress in chinese social media: a text mining and machine learning study. *Journal of Medical Internet Research*, 19(7):e243.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*.

John R Crawford and Julie D Henry. 2003. The depression anxiety stress scales (dass): Normative data and latent structure in a large non-clinical sample. *British journal of clinical psychology*, 42(2):111–131.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. 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*, pages 2098–2110. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jingcheng Du, Yaoyun Zhang, Jianhong Luo, Yuxi Jia, Qiang Wei, Cui Tao, and Hua Xu. 2018. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making*, 18(2):43.

Cecilia A Essau. 2005. Frequency and patterns of mental health services utilization among adolescents with anxiety and depressive disorders. *Depression and anxiety*, 22(3):130–137.

King-wa Fu, Ka Y Liu, and Paul SF Yip. 2007. Predictive validity of the chinese version of the adult suicidal ideation questionnaire: Psychometric properties and its short version. *Psychological Assessment*, 19(4):422.

Li Guan, Bibo Hao, Qijin Cheng, Paul SF Yip, and Tingshao Zhu. 2015. Identifying chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model. *JMIR mental health*, 2(2):e17.

Li Guan, Bibo Hao, and Tingshao Zhu. 2014. How did the suicide act and speak differently online? behavioral and linguistic features of china's suicide microblog users. *arXiv preprint arXiv:1407.0466*.

Keith M Harris and Melissa Ting-Ting Goh. 2017. Is suicide assessment harmful to participants? findings from a randomized controlled trial. *International journal of mental health nursing*, 26(2):181–190.

Keith M Harris, John P McLean, and Jeanie Sheffield. 2014. Suicidal and online: How do online behaviors inform us of this high-risk population? *Death studies*, 38(6):387–394.

Keith M Harris, Jia-Jia Syu, Owen D Lello, YL Eileen Chew, Christopher H Willcox, and Roger HM Ho. 2015. The abcs of suicide risk assessment: Applying a tripartite approach to individual evaluations. *PLoS One*, 10(6):e0127442.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Julie D Henry and John R Crawford. 2005. The short-form version of the depression anxiety stress scales (dass-21): Construct validity and normative data in a large non-clinical sample. *British journal of clinical psychology*, 44(2):227–239.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562.

Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. 2014. 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*, pages 844–849. IEEE.

Yen-Pei Huang, Tiong Goh, and Chern Li Liew. 2007. Hunting suicide notes in web 2.0-preliminary findings. In *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, pages 517–521. IEEE.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the us. *Crisis*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Marcel Adam Just, Lisa Pan, Vladimir L Cherkassky, Dana L McMakin, Christine Cha, Matthew K Nock, and David Brent. 2017. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature human behaviour*, 1(12):911.

Diederik Kingma and Jimmy Ba. 2015. ADAM: A method for stochastic optimization. In *Proc. of ICLR*.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*.

Tim MH Li, Ben CM Ng, Michael Chau, Paul WC Wong, and Paul SF Yip. 2013. Collective intelligence for suicide surveillance in web forums. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 29–37. Springer.

Meizhen Lv, Ang Li, Tianli Liu, and Tingshao Zhu. 2015. Creating a chinese suicide dictionary for identifying suicide risk on social media. *PeerJ*, 3:e1455.

Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide ideation of individuals in online social networks. *PloS one*, 8(4):e62262.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

World Health Organization et al. 2014. *Preventing suicide: A global imperative*. World Health Organization.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.

John P Pestian, Michael Sorter, Brian Connolly, Kevin Bretonnel Cohen, Cheryl McCullumsmith, Jeffry T Gee, Louis-Philippe Morency, Stefan Scherer, Lesley Rohlfs, and STM Research Group. 2017. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. *Suicide and Life-Threatening Behavior*, 47(1):112–121.

Fuji Ren, Xin Kang, and Changqin Quan. 2015. Examining accumulated emotional traits in suicide blogs with an emotion topic model. *IEEE journal of biomedical and health informatics*, 20(5):1384–1396.

Debra J Rickwood, Frank P Deane, and Coralie J Wilson. 2007. When and how do young people seek professional help for mental health problems? *Medical journal of Australia*, 187(S7):S35–S39.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 167–175.

Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat-Seng Chua, and Wendy Hall. 2018. Cross-domain depression detection via harvesting social media. In *IJCAI*, pages 1611–1617.

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

SPSS Tutorials. 2014. Pearson correlation. *Retrieved on February*.

M. Johnson Viouls, B. Moulahi, J. Az, and S. Bringay. 2018. Detection of suicide-related posts in twitter data streams. *Ibm Journal of Research & Development*, 62(1):7:1–7:12.

Henrik D Zachrisson, Kjetil Rödje, and Arnstein Mykletun. 2006. Utilization of health services in relation to mental health problems in adolescents: a population based survey. *BMC public health*, 6(1):34.

Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Tingshao Zhu. 2014a. Using linguistic features to estimate suicide probability of chinese microblog users. In *International Conference on Human Centered Computing*, pages 549–559. Springer.

Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Tingshao Zhu. 2014b. Using linguistic features to estimate suicide probability of chinese microblog users. In *International Conference on Human Centered Computing*, pages 549–559. Springer.