future internet

MDPI

Article

Incorporating Background Checks with Sentiment Analysis to Identify Violence Risky Chinese Microblogs †

Yun-Fei Jia [1,*], Shan Li [2] and Renbiao Wu [1]

[1] School of Electronic Engineering and Automation, Civil Aviation University of China, Tianjin 300300, China; rbwu@cauc.edu.cn

[2] Honeywell Technology Solutions China, Beijing 100015, China; shanzai0805@163.com

* Correspondence: yfjia@cauc.edu.cn

† This paper is an extended version of our paper published in the IEEE 17th International Conference on Information Reuse and Integration, Pittsburgh, PA, USA, 28–30 July 2016; pp. 463–468.

Received: 26 July 2019; Accepted: 5 September 2019; Published: 19 September 2019

Abstract: Based on Web 2.0 technology, more and more people tend to express their attitude or opinions on the Internet. Radical ideas, rumors, terrorism, or violent contents are also propagated on the Internet, causing several incidents of social panic every year in China. In fact, most of this content comprises joking or emotional catharsis. To detect this with conventional techniques usually incurs a large false alarm rate. To address this problem, this paper introduces a technique that combines sentiment analysis with background checks. State-of-the-art sentiment analysis usually depends on training datasets in a specific topic area. Unfortunately, for some domains, such as violence risk speech detection, there is no definitive training data. In particular, topic-independent sentiment analysis of short Chinese text has been rarely reported in the literature. In this paper, the violence risk of the Chinese microblogs is calculated from multiple perspectives. First, a lexicon-based method is used to retrieve violence-related microblogs, and then a similarity-based method is used to extract sentiment words. Semantic rules and emoticons are employed to obtain the sentiment polarity and sentiment strength of short texts. Second, the activity risk is calculated based on the characteristics of part of speech (PoS) sequence and by semantic rules, and then a threshold is set to capture the key users. Finally, the risk is confirmed by historical speeches and the opinions of the friend-circle of the key users. The experimental results show that the proposed approach outperforms the support vector machine (SVM) method on a topic-independent corpus and can effectively reduce the false alarm rate.

Keywords: sentiment analysis; violence risk; topic independent; semantic similarity; semantic rules

1. Introduction

With the rapid development of Web 2.0, more and more people retrieve and share information on social media. Microblog, as a popular user-generated content service, is attracting more and more users. Unlike conventional web text, its length is limited to 140 characters. This feature heightens user engagement in publishing their opinions more frequently and quickly. Microblogging has grown into a rich corpus of users' emotions and opinions, especially on hot topics. Consequently, sentiment analysis of microblog text has attracted much attention in recent years and has spread from computer science to management sciences and social sciences due to its importance in business and society [1]. There is an increasing interest in sentiment analysis of microblogs in several areas, such as predicting consumer reviews [2], market trends [3] and specific events [4]. Most of the works provided in the literature depend on specific training data. They usually perform well only when there is a good match between the training and test data. In order to train an accurate domain-specific sentiment classifier, a large

number of labeled samples are needed, which are costly and time-consuming to obtain [5]. Our previous work illustrated a semantic rule-based topic-independent sentiment analysis approach [6]. This paper is an expanded edition of [6] and combines the approach in [6] with background checking to detect violent Chinese microblogs.

Violence risk assessment of microblogs has been rarely reported in the literature. This is contradictory to the observation that more and more terrorism speeches and violence threats are published on the Internet every day. This phenomenon can be attributed to several reasons: (1) there is no definitive annotated corpus, which is important for machine learning-based studies; (2) most violent microblogs are insincere or a form of emotional catharsis, which hinder such studies; (3) the violence risk assessment of microblogs involves multiple topics, which is not well-addressed by the existing topic-dependent methods. In recent years, the topic-independent textual analysis has attracted increasing attention. For English corpuses, Martineau et al. [7] introduced a method to determine topic-independent bias scores for words. Read and Carroll [8] proposed weakly supervised techniques to reduce the dependency on training data. The topic-independent sentiment analysis for Chinese text is still rarely reported, probably because the characteristics of Chinese text make it more difficult to analyze: (1) A Chinese microblog contains more information than English within the same length limit, while the syntax and semantics are more diverse; (2) There are usually several subsentences contained in one message, and they may not necessarily have coherent sentiment [9]. That is to say, it is very likely to have two opposite emotions within one microblog [10]; (3) Unlike English, a Chinese sentence has no spaces between words. Each Chinese sentence should be first split into a sequence of words before further analysis. Therefore, the methods of analysis of English corpuses cannot be directly applied to Chinese text.

This paper aims to detect violent Chinese microblogs with a lower false alarm rate. Hence, the approach should not depend on topic-specific training data. The contributions of this paper are summarized as follows:

1. A framework for step by step evaluation of the violence risk of Chinese microblogs is proposed.
2. Sentimental analysis is combined with background checks for violence activity detection.
3. Our method is topic-independent and easy to apply.

This paper is arranged as follows: Section 2 describes related works. Section 3 describes our approach for the calculation of polarity of sentiment and estimation of risk. Experimental design and validation are included in Section 4. Section 5 concludes this paper.

2. Related Work

Current reports on sentiment classification can be roughly divided into three parts: sentiment-knowledge-based [11], conventional machine learning-based [12], and deep learning-based [1]. The knowledge-based approach utilizes the sentiment lexicon, rules, and patterns. Its accuracy depends on the maturity of the lexicons. Machine learning approaches normally use the bag of words, part of speech, and so on, as features to train the mathematical models. The most commonly used conventional machine learning models include support vector machine (SVM), naive Bayes (NB), and maximum entropy (ME). Deep learning approaches usually extract text features automatically and train classifiers through deep learning models such as convolutional neural network (CNN) and recurrent neural network (RNN).

Sentiment-knowledge-based approaches use universal or domain-specific dictionaries to obtain emotion. Xu et al. [13] built an extended sentiment dictionary including basic sentiment words, field sentiment words, and polysemic sentiment words. Their approach is effective for sentiment recognition in Chinese comment texts. Tang et al. [14] built a large-scale sentiment lexicon by a representation approach and treated sentiment lexicon learning as a phrase-level sentiment classification. Hamidreza et al. [15] built adaptive sentiment lexicons to improve the accuracy of polarity classification in microblogs.

The approaches based on machine learning demand a number of meaningful well-adopted features as input, and then uses various classifiers to determine the sentiment polarity. Pang et al. [12] applied machine learning to sentiment classification for the first time, tried two text features of *n*-gram and part of speech, and found that SVM had the best accuracy. Xie et al. [16] proposed a hierarchical structure-based hybrid approach for sentiment analysis and analyzed the contribution of various features in SVM. Tan et al. [17] presented a semisupervised and user-level framework to predict the sentiment polarity by utilizing information about user–user relationships in social media. Socher et al. [18] proposed recursive deep models to compute compositional representations for variable length and syntactic type phrases. Those models are employed as features to classify each phrase.

Approaches based on deep learning usually learn from large-scale labeled data to high-level feature representations of text through a multilayered neural network. Kim et al. [19] used a CNN to incorporate much more parameters to accurately classify the sentiment polarity of movie reviews. However, CNNs struggle to capture the context in a piece of a message, particularly for a Chinese microblog, which includes much more information than an English counterpart. Wang et al. [20] applied a long short-term memory (LSTM) network for twitter sentiment prediction. This type of deep neural network-based model employs gated units to select appropriate parameters for time series. Yang et al. proposed a HHAS model to improve the accuracy of aspect-level sentiment analysis [21].

The above methods usually demand large amounts of well-labeled data. To some extent, the performance of deep learning methods depends on the quality of the annotated corpus. For example, at least 10,000 pieces of messages are required to train the deep neural networks. Unfortunately, domain-specific datasets cannot be obtained easily. Moreover, because social media contain a large range of topics, labeling the data manually will inevitably incur errors [22]. This problem has attracted the attention of researchers in recent years. Semisupervised classification significantly reduces the demand of labeled data. The variational autoencoder (VAE) has been widely used to automatically annotate unlabeled data. However, this approach is immature and cannot be directly used in practice [23].

To sum up, the deep learning method can output good results based on sufficiently large well-labeled datasets but performs poorly on a small sample that includes a large range of topics. Moreover, the semisupervision method is still immature and cannot be used in practice. In the case of violence-prone Chinese microblog detection, the volume of the corpus is small, and there are diverse topics in the datasets. Therefore, a semantic rule-based method can still be effective.

3. Our Approach

Because our aim is to detect violent speeches on the Internet with a lower false alarm rate, we have two main tasks: (1) calculation of activity risk of the target microblog; and (2) investigation of the background of the microblog user to identify if he/she has a crime tendency. Consequently, our approach can be described as two stages: (1) calculation of activity risk; and (2) background check. This is illustrated in Figure 1.

As illustrated in Figure 1, at the first stage, a sentiment analysis method based on dictionary and rules are employed to identify the speeches with potential violence risk. It can be briefly described as follows: We extract sentiment words from Chinese microblogs, and then employ semantic rules and emoticons to calculate the sentiment polarity. In the following, the activity risk is calculated based on the semantic rules. In such a way, the potential violence publishers can be identified and treated as "key users". This part is described in Section 3.1. At the second stage, a comprehensive judgment is carried out to reduce the false alarm rate, which will be detailed in Section 3.2. The philosophy behind this is to further determine whether the key users have long-term negative sentiments or unhealthy friendships by using their historical speeches and opinions or attitude from his/her circle of friends. After all, one will be unlikely to commit crime arbitrarily if he/she has long-lasting positive sentiment and good friendships [24].

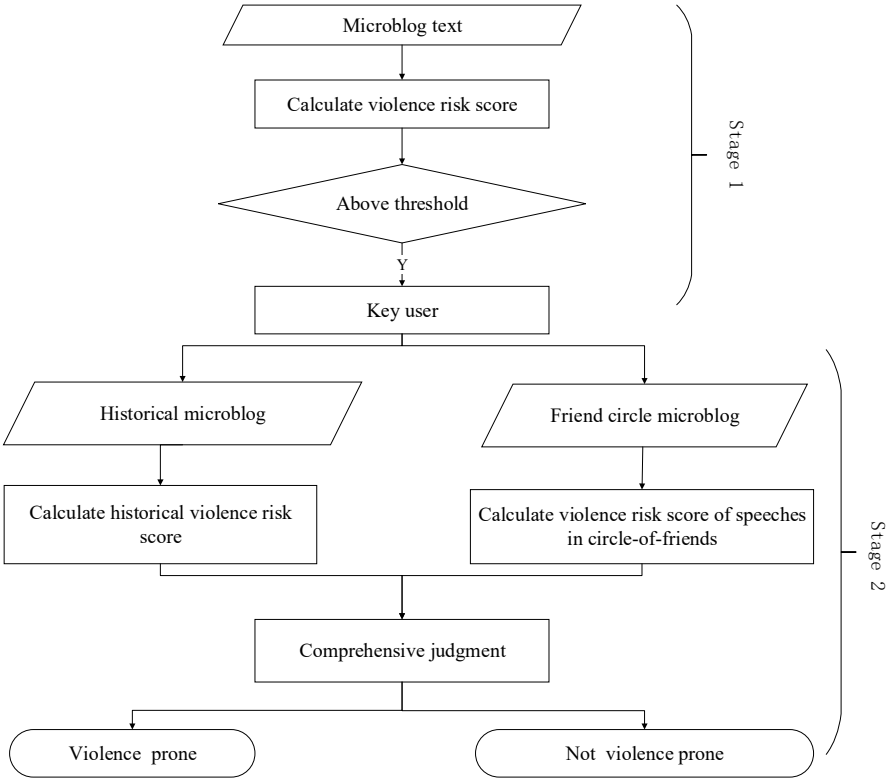

Figure 1. Calculation of comprehensive violence risk.

3.1. Calculation of Violence Risk Score

This section focuses on the violence risk score of each microblog text through a sentiment analysis algorithm. Microblog users having high violence risk scores will be treated as key users, i.e., potentially risky users. In Section 3.2, the key users will be subject to background checks to reduce the false alarm rate.

3.1.1. Extraction of Subjective Microblogs

In microblogs, there are many objective news or reports which are forwarded by some users. Since only subjective speeches can be treated as threatening, it is necessary to eliminate objective speeches or news mixed in the microblogs.

Some radical or violence-prone messages could incur social panic even if they are not serious or emotionally cathartic. By longtime observations, more often than not, such messages include the first-person words such as "I", "me", or "we". It should be clearly pointed out that our approach is not targeted for detecting secret signals or languages that are used by true terrorists. Another feature of subjective opinions is that the relevant microblog is usually forwarded less than three times. In this paper, we use both features to eliminate objective news.

3.1.2. Extraction of Sentiment Words

Firstly, we need to determine the words associated with emotional tendency in the microblog text. Because new expressions or words have been emerging on the Internet, the existing lexicon has to be continuously complemented. In this paper, we use two methods to extend the existing lexicon, i.e., manually annotating new sentiment words, and use semantic similarity to estimate sentiment polarity of sentiment words. Both these methods are detailed as follows:

(1) Manually annotating new sentiment words: the sentiment words have two typical properties: polarity and strength. The sentiment vocabulary ontology provided by Dalian University of

Technology [25] includes 27,466 entries, in which each has a polarity value of 0, 1, or 2, representing neutral, positive, and negative, respectively, and each has one of five strength levels, i.e., 1, 3, 5, 7, 9 (the strength increases in this order). We utilize it as sentiment lexicon since it covers the majority of common expressive words. As a kind of social media, microblog text has informal and colloquial characteristics. The frequency of network terms is very high, and they play an important role in emotion detection, but these terms are not included in a typical sentiment lexicon. Thus, it is necessary to complement network words using a sentiment lexicon. In the "wangci" Website [26], there are popular network vocabularies and their explanation. In this paper, 494 items are manually annotated and added to our lexicon according to their polarity and strength. We set the strength value according to the same standard of sentiment ontology in [24].

(2) Semantic similarity: Some new words in a microblog may be excluded from the existing vocabulary ontology. We use a semantic similarity algorithm to extract the sentiment words. Only some specific parts of speech can be selected as sentiment words, we hence selected nouns, verbs, and adjectives for further analysis. Our semantic similarity algorithm refers to HowNet [27], which works well for measuring word similarity. It can be briefly described as follows: for two Chinese word items w_1 and w_2, if w_1 has n senses (concepts): $x_1, x_2, \ldots, x_n$, and w_2 has m senses: $y_1, y_2, \ldots, y_m$, the similarity between w_1 and w_2 is the maximum similarity of each sense pair, which can be formulated as:

$$Sim(w_1, w_2) = \max_{i \in [1,n], j \in [1,m]} (Sim(x_i, y_j)) \tag{1}$$

where the similarity of the two senses can be calculated as follows:

$$Sim\left(x_i, y_j\right) = \frac{\lambda}{\lambda + d(x_1, y_2)} \tag{2}$$

In Equation (2), λ denotes a variable positive number, $d(x_1, y_2)$ refers to the distance between the sense x_1 and y_2 in the hierarchal tree. The range of $Sim(x_1, y_2)$ lies in [0, 1]. The value of λ and the computation of $d(x_1, y_2)$ refers to HowNet.

For any word, the emotional value can be calculated by the distance between the word and base words. The principle is to compute the distance between the specific word w and every word item in positive/negative base words, respectively. At last, we get the word sentiment value through the comparison between their mean values. This can be formulated as follows:

$$S_w = \frac{1}{n} \sum_{i=1}^{n} Sim(w, p_i) - \frac{1}{m} \sum_{j=1}^{m} Sim(w, n_j) \tag{3}$$

where p_i denotes one of the positive base words, n denotes the number of positive base words. Similarly, n_j denotes one of the negative base words, and m denotes the number of negative based words. Consequently, the calculated S_w falls into $(-1, 1)$. Because words whose emotional value reaches a certain extent can be considered sentiment words, a threshold must be set to filter out those with a weak emotional value. Accordingly, we set a threshold T to determine whether or not it can be treated as a sentiment word. For example, the word w is judged to be an emotional word if $|S_w| > T$. The strength of the sentiment word w is $|S_w|$. To ensure the same scale of sentiment strength as the sentiment words in the lexicon, we scale the strength values of the lexicon from 1, 3, 5, 7, 9 to 0.1, 0.3, 0.5, 0.7, 0.9.

3.1.3. Semantic Rules

In this section, the text scoring rules are introduced. Due to the complexity of the Chinese syntax, the sentiment words have various forms of expression. Since emotional words have been extracted and assessed, we can calculate the score of subsentences using semantic rules. The sentiment score of a simple sentence can be calculated by the intrasentence rules. As for the overall microblog, which may include several sentences, the relationship between two sentences should also be considered by intersentence rules. According to different expressions of semantic rules, an essential parameter

z_n, $n \in (1, 2, 3, 4)$ is introduced to enhance, weaken, or reverse the sentiment score of a sentence. More specifically, z_1 can be used to weaken the sentiment score for a specific expression form by intrasentential rules, and z_2, z_3, and z_4 can be used to reverse, weaken, or enhance the sentiment score by intersentential rules.

Intrasentential rules: The inverse words can reverse the polarity of sentiment words, and the degree of adverbs can influence the intensity [28]. Hence, we collect these two kinds of words to make a corresponding dictionary. The reverse dictionary contains 26 commonly used items, such as "没有 (not)", "不 (no)", "未必 (without)", and so on. The degree adverb dictionary is collected from HowNet, and it is divided into six levels according to different strength. More specifically, the effect of adverbs can enhance or weaken the polarity of sentiment. If adverbs will enhance sentiment polarity, the value of adverbs can be set to be slightly larger than 1. Otherwise, it can be set to be slightly less than 1. This scoring rule is also applicable for other parameters described below, such as z_n, $n \in (1, 2, 3, 4)$. Table 1 shows some of the degree adverb words and their strength.

Table 1. Degree adverb word and strength.

Degree	Example Words	Number	Strength
Most	不得了 (extremely), 绝对 (absolutely)	69	2
Over	出头 (a little over), 过度 (excessively)	30	1.7
Very	不过 (moderately), 颇为 (mildly)	42	1.5
More	大不了 (at the worst), 更加 (all the more)	37	1.3
-ish	少许 (a little), 未免 (a bit too)	29	0.8
Insufficiently	相对 (relatively), 丝毫 (in the least)	12	0.5

The following shows four intrasentential semantic rules and the corresponding phrase combination score s_s. Let n be the number of the inverse word, p_s be the strength of the sentiment word, p_p be the polarity of the sentiment words, and a_s be the strength of the degree adverb.

(a) Degree adverb + sentiment word:

$$s_s = a_s \times p_s \times p_p \tag{4}$$

(b) Inverse + sentiment word: The inverse words will change the polarity of a phrase according to the number before it.

$$s_s = (-1)^n \times p_s \times p_p \tag{5}$$

where n refers to the number of inverse words in the sentence.

(c) Degree adverb + inverse + sentiment word: This kind of phrase only switches the polarity of the sentences.

$$s_s = (-1) \times a_s \times p_s \times p_p \tag{6}$$

(d) Inverse + degree adverb + sentiment word: The inverse word is in the foremost position of the phrase, leading to the weakness of the sentiment strength. We hence set the value $z_1 = 0.5$.

$$s_s = (-1) \times s_s \times p_s \times p_p \times z_1 \tag{7}$$

Intersentential rules: If a microblog message includes several simple sentences, these sentences are usually not independent, so summing up the values directly without considering the correlation between them will lead to misclassification [29]. As conjunction words connect two simple sentences, we create a model for different types of conjunctions as follows:

(1) Transition relationship. The polarity of the later sentence will reverse, and the whole sentence polarity is consistent with the latter.

$$\text{Sen} = z_2 \times Sen_1 + Sen_2, \quad z_2 = -1 \tag{8}$$

(2) Progressive relationship. The strength of the whole sentence will be enhanced.

$$\text{Sen} = z_3(Sen_1 + Sen_2), \quad z_3 = 1.5 \tag{9}$$

(3) Concession relationship. The polarity of the later sentence will reverse, and the final sentence polarity is the same as the former.

$$\text{Sen} = Sen_1 + z_4 \times Sen_2, \quad z_4 = -1 \tag{10}$$

Combining the intrasentential and intersentential rules, the sentiment score of the Weibo text can be calculated by:

$$score_{text} = \frac{1}{n}\sum\nolimits_{i=1}^{n} s_n + \frac{1}{m}\sum\nolimits_{(j=1)}^{m} Sen_m \tag{11}$$

where n denotes the number of sentiment words in an individual sentence, and m denotes the number of sentences in a microblog.

3.1.4. Contribution of Emoticons

One way to overcome the domain and topic dependency of short text is to use emoticons as an additional discriminant condition for sentiment scoring. People use emoticons frequently to express their opinions in social media. Since the semantic rules library cannot cover all expression patterns, the emoticons implicate the sentiment to a large extent [30]. Therefore, it is reasonable to incorporate the sentiment score of emoticons into our approach.

We select the built-in emoticons of Sina Weibo [31], a popular Chinese microblog, and classify them into positive and negative categories. The emoticons are usually considered to have equivalent impact with sentiment words. We hence set the strength of emoticons into (−1, 1), i.e., positive strength refers to positive sentiment, and vice versa. Table 2 shows some of the emoticons and their strength.

Table 2. Emoticon examples.

Positive	Negative	Strength
		±1
		±0.8
		±0.6
		±0.4
		±0.2

For each microblog, we can simply calculate the emoticon value by retrieving the number and strength of the emoticons:

$$score_{emo} = \frac{1}{N}\sum\nolimits_{(i=1)}^{N} e_i \tag{12}$$

where e_i denotes the designated strength value of the ith emoticon, and N denotes the total number of emoticons.

3.1.5. Final Score

The final score of violence risk can be calculated as weighted average of the score of text and score of emoticons. It can be formulated as follows:

$$s_1 = \propto \times score_{emo} + \beta \times score_{text} \tag{13}$$

where both parameters $\propto$ and β are in [0, 1), and $\propto + \beta = 1$. Consequently, s_1 falls into $(-1, 1)$.

3.1.6. Activity Risk

The activity risk refers to whether or not the microblogs implicate implementation of violence. This can be determined via Chinese semantic rules. In the following, we can determine the violence risk of microblogs via the following expression:

$$\mathrm{D} = \frac{1}{2}(s_1 - s_2 < w_1,\ w_2 >) \tag{14}$$

where D denotes the violence risk, s_1 refers to the sentiment which is calculated by Equation (13). The range of s_1 is limited in $(-1, 1)$. $s_2 < w_1,\ w_2 >$ refers to the activity risk of microblogs. Its value depends on the activity words and location words included in the microblog. For example, a negative s_1 means the microblog user has discontent emotions, and a positive s_2 implies he/she will be acting. That is to say, a larger negative D tells us the user has larger violence risk. w_1 refers to location words, and w_2 denote activity words. The property of w_2 can usually be divided into a direct type or an indirect type. If w_2 is a direct type verb, $s_2 < w_1,\ w_2 >$ adopts the strength of that word. Otherwise, $s_2 < w_1,\ w_2 >$ depends on whether or not w_1 exists. Under this condition, if both w_1 and w_2 exists, $s_2 < w_1,\ w_2 >$ adopts the strength of w_2. Otherwise, $s_2 < w_1,\ w_2 >$ is treated as 0.

3.2. Background Checks of Key Users

In this paper, background check refers to sentiment analysis of historical microblogs of the key users and relevant opinions published by their internet friends (or circle of friends). The historical microblogs can tell us whether the key user has long-lasting radical ideas, and the opinions of his/her circle of friends will galvanize or dissuade the key user. The background check is carried out in two steps: (1) calculation of the sentiment of historical microblogs; and (2) scale the results in step 1 by the friends' opinions.

3.2.1. Sentiment of Historical Microblogs

A person having a penchant for crime usually exhibits some signs beforehand. A typical sign is longtime negative sentiment. This can be determined via in-depth exploration of the personal details and historical microblogs of these key users. Usually, the closer the release time, the more impact on current sentiment of the key user. Accordingly, the historical sentiment score of the key user can be approximated by

$$s_{history} = (\sum_{i=1}^{n} s_i \frac{1}{t})/n \tag{15}$$

Its value depends on the violence risk score of each historical microblog and the time of publication. $i \in [1, \mathrm{n}]$ refers to the number of all historical microblogs which exhibit obvious positive/negative sentiment. s_i refers to the violence risk score of each historical microblog. t refers to the number of days from the publishing time of historical microblog to now. A negative value of $s_{history}$ can roughly indicate that the key user has long-lasting negative sentiment.

3.2.2. Opinion of the Key User's Circle of Friends

Whether or not a radical idea is supported or opposed by his/her friends could be an important factor to galvanize or dissuade the actor. In the case of microblogs, each microblog user has several followers. Similarly, he/she can follow other microblog users. In such a way, the followers of the key users can immediately see what is issued by the key user, and comment on it. The user and his/her followers interact with each other through thumb-up, @, or comment, forming a circle of friends. Once a key user is identified via the aforementioned steps, our crawler will immediately follow this key user and find his/her circle of friends in his/her homepage of the microblog. It should be clearly pointed out that these comments, thumb-up, or @ are nonprivate and can be retrieved by our crawler.

The possibility of violent action depends on the opinions of the circle-of-friends to some extent. The closer the relationships the friend has, the more impact they have on the key user's actions. For the first step, multiple features are used to calculate the closeness coefficients σ for the friends of the key user.

$$\sigma_j = \sum_{i=1}^{n}(A_{ji} + B_{ji} + C_{ji}) / \sum_{j=1}^{m}\sum_{i=1}^{n}(A_{ji} + B_{ji} + C_{ji}) \tag{16}$$

where A_{ji}, B_{ji}, and C_{ji} refer to the total number of thumb-ups, @, and comments, respectively, from the jth friend to the ith historical microblog of the key user. n refers to the total number of the key user's historical microblogs, and m refers to the total number of the key user's friends. A larger σ means a closer relationship between the user and his/her friend. Consequently, Equation (15) can be expanded as:

$$o_{fi} = \sum_{j=1}^{m} \sigma_j p_{ji} \tag{17}$$

o_{fi} refers to the overall opinion of all friends to the ith historical microblog of key user. p_{ji} refers to the polarity of opinions, i.e., supporting or opposing, of the jth friend to the ith historical microblog. m refers to the total number of key user's friends. Thus, Equation (15) can be expanded to:

$$s_{bg} = (\sum_{i=1}^{n}\sum_{j=1}^{m} \sigma_j p_{ji} s_i \frac{1}{t}) / n \tag{18}$$

where s_{bg} refers to the result of background checks. Equation (18) indicates that the final sentiment score of each historical microblog will be affected by friends' opinions to some extent. Accordingly, a threshold can be set to determine whether the key user will be warned. Similarly, a negative value of s_{bg} can roughly indicate that the key user has a long-lasting negative sentiment.

4. Experiment and Results

4.1. Datasets and Criteria

Because there are no existing definitive Chinese datasets that are randomly selected and annotated, we compiled a crawler tool based on the scrapy framework [32]. This tool can collect data from Sina Weibo. To ensure topic-independence of the dataset, we randomly selected target users and retrieved the first five pages of their microblogs as experimental data. Ten volunteers were employed to manually annotate the polarity of each microblog. In order to eliminate arbitrary results, the ten volunteers had different ages, genders, and professional background, and the final results were averaged over the annotation of the ten volunteers. To keep balance of the corpus, the numbers of positive and negative training samples were both 5000, and the number of test data was 1000.

We take the precision (denoted by p_{pos} and p_{neg}), recall (denoted by R_{pos} and R_{neg}), and F1 score (denoted by $\mathrm{F1}_{pos}$ and $\mathrm{F1}_{neg}$) as the evaluation criteria. We describe TP as the number of correctly classified positive samples, TN as the number of correctly classified negative samples, FP as the number of falsely classified positive samples, and FN as the number of falsely classified negative samples. The criteria are expressed as follows:

$$\mathrm{p}_{pos} = \frac{TP}{TP + FP} \tag{19}$$

$$\mathrm{p}_{neg} = \frac{TN}{TN + FN} \tag{20}$$

$$\mathrm{R}_{pos} = \frac{TP}{TP + FN} \tag{21}$$

$$\mathrm{R}_{neg} = \frac{TN}{TN + FP} \tag{22}$$

$$F1_{pos} = \frac{2P_{pos} \times R_{pos}}{P_{pos} + R_{pos}} \quad (23)$$

$$F1_{neg} = \frac{2P_{neg} \times R_{neg}}{P_{neg} + R_{neg}} \quad (24)$$

4.2. Preprocessing

Unlike English and other European languages, Chinese texts have no spaces between words. Therefore, word segmentation is an important procedure in Chinese natural language processing. The Chinese Lexical Analysis System (ICTCLAS) provided by the Institute of Computing Technology is a widely used Chinese word segmentation tool [33]. Firstly, we use punctuation to divide a piece of microblog into several subsentences. Then, useless compositions such as URLs, hashtags, @, and usernames within the microblog are removed. Finally, each word is segmented and extracted.

4.3. Parameter Setting

When utilizing the semantic similarity method to extract sentiment words, the base positive/negative words should be carefully selected to calculate the polarity of a word. Therefore, we extract the most frequent words in the lexicon from the experimental data. Recall that in Section 3.1, we mentioned that the number of base words and a threshold T should be determined. This is a trial and error process. We selected 10, 20, and 30 pairs of positive/negative words and set the threshold values as 0.3, 0.4, and 0.5. Thus, we obtained nine sets of classification results. Table 3 shows the accuracy of classification without considering the emoticons, from which we can clearly see the number of words should be set to 30 and the value T should be set to 0.4.

Table 3. Parameter setting.

	T = 0.3	**T = 0.4**	**T = 0.5**
N = 10	64.3	65.7	64.8
N = 20	65.7	67.2	66.5
N = 30	65.9	68.5	67.1

In addition, $\propto$ refers to the weight of emoticon, β refers to the weight of text, and $\propto + \beta = 1$ (see Section 3.1). In order to extract the feature of Chinese microblogs, we conducted several experiments on our training datasets to get the optimal results, i.e., $\propto = 0.7$ and $\beta = 0.3$.

4.4. Calculation of Sentiment Polarity

In this section, we performed several comparisons to validate the effectiveness of our approach. Recall that our target is to detect microblogs with risk of violence. Under this condition, SVM is a relatively appropriate classifier due to its good performance for classification of small samples. The results of different methods on the test corpus are listed in Table 4. For the SVM classifier, we selected sentiment words in the lexicon, numbers of each PoS, and emoticons as features to train the sample. Both polarity and number of emoticons were considered in the features. The parameters of the SVM classifier were selected in the training data by 10-fold cross-validation.

In addition, a CNN model was also used for comparison. The CNN model has 128 convolutional units, with a context window of three. The number of hidden units is 300. The Adam optimization algorithm is used to update the parameters at the training stage, and features are extracted via a mean pooling algorithm.

Table 4. Comparisons of different methods' outcomes.

	Positive			Negative		
	p_{pos}	R_{pos}	$F1_{pos}$	p_{neg}	R_{neg}	$F1_{neg}$
dic	62.5	69.4	65.7	65.6	58.3	61.7
sim	62.7	68.7	65.6	65.4	59.2	62.2
dic + sim	68.7	76.6	72.4	73.5	65.1	69.1
dic + sim + emo	69.7	77.7	73.5	74.8	66.2	70.2
SVM	68.5	76.3	72.2	73	64.1	68.2
CNN	67.9	75.4	71.8	71.5	63.2	67.6

The first four methods utilize the semantic rules to calculate the final score because it is the criterion to decide the sentiment tendency of a text. Firstly, we compared three different methods of sentiment words selection, i.e., the dictionary method (dic), similarity method (sim), and a combination of the two (dic + sim). From the results, we can see that the combined method performs best. Then, taking the emoticons into consideration, the proposed approach of this paper is expressed by dic + sim + emo, which outputs the best results. Finally, compared with the machine learning methods, SVM and CNN, the accuracy of our approach is improved in both positive and negative corpuses. The training accuracy of the CNN is higher, but its predicative accuracy is lower than that of SVM and that of our approach. This can be accounted for by the overfitting at the training stage on our small datasets. The results show that more features can effectively improve sentiment classification performance in topic-independent datasets.

In the next step, the violence- or terrorism-related words in the negative sentiment microblogs were detected. The resulting microblogs will be treated as key users for further validation.

4.5. Calculation of Activity Risk

Following the calculation method proposed in Section 3.2, the historical microblogs and circle-of-friends will be subject to further sentiment analysis. The last 180 days of historical microblogs are investigated to determine whether the key user has long-lasting negative sentiment. More specifically, the key users with long-term positive sentiment or healthy friendships are be removed to reduce the false alarm rate, because they may publish a joking or cathartic microblog and do not tend to act on the setiment. Finally, we get the results shown in Table 5.

Table 5. Calculated threat risk.

Microblogs	Violence	Threat Level
发个炸弹，把国航的航班都炸飞，什么态度总让我老公误机哼 (Sending a bomb to Air China's flights, what attitude, always makes my husband miss the opportunity)	−1.0	High
飞行时间好久啊，我想在飞机上抽烟！ (Flying time is long, I want to smoke on the plane!)	−0.5	Middle
买个机票，航空公司服务人员态度好差，好想冲到机场讨个说法！😤 (Buying a ticket, the airline service staff is in a bad attitude, I really want to rush to the airport to discuss! 😤)	−0.12	Low

The value of D is calculated using Equation (14), and a negative value indicates a threat. When the value of D approaches −1, the threat is high. We hence roughly divide the section of [−1, 0] into three parts, $-0.7 \geq D \geq -1.0$ implicates a high threat, $-0.3 \geq D > -0.7$ indicates a moderate threat, and $0 \geq D > -0.3$ means low threat.

5. Conclusions

More and more violent threats are appearing on the Internet, especially through social media such as Chinese microblogs. Such microblogs usually have strong sentiment polarity and yet have not been a central topic of research. Conventional topic-dependent semantic methods or machine learning methods cannot address this topic. Moreover, topic-independent sentiment analysis of Chinese short text is rarely reported in the literature. In this paper, we propose a framework to comprehensively assess the violence risk of microblog users. First, we assess the polarity and strength of emerging networks of words by using a similarity-based method, based on which we complement the existing lexicon. Second, Chinese semantic rules combined with emoticons are employed to assess the activity risk. Finally, the filtered key users will be subject to further background checks, i.e., sentiment analysis of historical microblogs and opinions of his/her circle of friends. This can validate whether the user has long-term negative sentiment and healthy friendships. Experimental results show that our approach can output good binary sentiment classification compared with SVM method, and hence our method can effectively detect microblogs of high violence risk. In addition, our approach does not require a training corpus, is easy to implement, and can be generalized to analyze other types of Chinese short text.

Author Contributions: Conceptualization, Y.-F.J. and S.L.; methodology, Y.-F.J.; software, Y.-F.J. and S.L.; validation, R.W.; formal analysis, Y.-F.J. and S.L.; investigation, S.L.; resources, R.W. and S.L.; data curation, S.L.; writing—original draft preparation, Y.-F.J.; writing—review and editing, Y.-F.J.; supervision, R.W.; project administration, R.W.; and funding acquisition, R.W.

Funding: This research was funded by [National Key Research and Development Program of China] grant number [2018YFC0823701], [Natural Science Foundation of Tianjin] grant number [19JCYBJC15900] and [the Open Fund of Tianjin Key Lab for Advanced Signal Processing] grant number [2017ASP-TJ04].

Acknowledgments: We thank anonymous reviewers and editors for evaluable comments on earlier drafts of the manuscript, and ZhiQuan Zhou for his helpful comments on writing.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1253.
2. Fang, Y.; Tan, H.; Zhang, J. Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness. *IEEE Access* **2018**, *6*, 20625–20631. [CrossRef]
3. Ren, R.; Wu, D.D.; Liu, T.X. Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine. *IEEE Syst. J.* **2019**, *13*, 760–770. [CrossRef]
4. Aloufi, S.; Saddik, A.E. Sentiment Identification in Football-Specific Tweets. *IEEE Access* **2018**, *6*, 78609–78621. [CrossRef]
5. Yuan, Z.; Wu, S.; Wu, F.; Liu, J.; Huang, Y. Domain Attention Model for Multi-Domain Sentiment Classification. *Knowl. Based Syst.* **2018**, *155*, 1–10. [CrossRef]
6. Han, P.; Li, S.; Jia, Y.F. A Topic-Independent Hybrid Approach for Sentiment Analysis of Chinese Microblog. In Proceedings of the IEEE 17th International Conference on Information Reuse and Integration, Pittsburgh, PA, USA, 28–30 July 2016; pp. 463–468.
7. Martineau, J.C.; Cheng, D.; Finin, T. Tisa: Topic independence scoring algorithm. In *Machine Learning and Data Mining in Pattern Recognition*; Springer: Berlin, Germany, 2013; pp. 555–570.
8. Read, J.; Carroll, J. Weakly supervised techniques for domain-independent sentiment classification. In Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, Hong Kong, China, 6 November 2009; pp. 45–52.
9. Cui, A.; Zhang, H.; Liu, Y.; Zhang, M.; Ma, S. Lexicon-Based Sentiment Analysis on Topical Chinese microblog messages. In *Semantic Web and Web Science*; Springer: Berlin, Germany, 2013; pp. 333–344.
10. Yan, B.; Yecies, B.; Zhou, Z.Q. Metamorphic relations for data validation: A case study of translated text messages. In Proceedings of the IEEE/ACM 4th International Workshop on Metamorphic Testing (MET '19), Montreal, QC, Canada, 26 May 2019; pp. 70–75.

11. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. Computational linguistics. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]
12. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA, 6–7 July 2002; pp. 79–86.
13. Xu, G.X.; Yu, Z.H.; Yao, H.S.; Li, F.; Meng, Y.; Wu, X. Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary. *IEEE Access* **2019**, *7*, 43749–43762. [CrossRef]
14. Tang, D.; Wei, F.; Qin, B.; Zhou, M.; Liu, T. Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In Proceedings of the COLING 2014: 25th International Conference on Computational Linguistics, Dublin, Ireland, 23–29 August 2014; pp. 172–182.
15. Keshavarz, H.; Abadeh, M.S. ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowl. Based Syst.* **2017**, *122*, 1–16. [CrossRef]
16. Xie, L.; Zhou, M.; Sun, M. Hierarchical structure based hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction. *J. Chin. Inf. Process.* **2012**, *26*, 73–83.
17. Tan, C.; Lee, L.; Tang, J.; Jiang, L.; Zhou, M.; Li, P. User-level sentiment analysis incorporating social networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1397–1405.
18. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.; Potts, C. Recursive deep models for semantic Compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
19. Kim, Y. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1746–1751.
20. Wang, X.; Liu, Y.; Sun, C.; Wang, B.X.; Wang, X.L. Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 1343–1353.
21. Yang, M.; Jiang, Q.N.; Shen, Y.; Wu, Q.Y.; Zhao, Z.; Zhou, W. Hierarchical human-like strategy for aspect-level sentiment classification with sentiment linguistic knowledge and reinforcement learning. *Neural Netw.* **2019**, *117*, 240–248. [CrossRef] [PubMed]
22. Xia, R.; Jiang, J.; He, H.H. Distantly Supervised Lifelong Learning for Large-Scale Social Media Sentiment Analysis. *IEEE Trans. Affect. Comput.* **2017**, *8*, 480–491. [CrossRef]
23. Xu, W.; Tan, Y. Semisupervised Text Classification by Variational Autoencoder. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, 1–14. [CrossRef] [PubMed]
24. Melick, M.D. The relationship between crime and unemployment. *Park Place Econ.* **2003**, *11*, 13.
25. Emotional Lexical Ontology Library. Available online: http://ir.dlut.edu.cn/EmotionOntologyDownload (accessed on 11 November 2018).
26. China's Internet Language. Available online: http://wangci.net/ (accessed on 17 November 2018).
27. Zhu, Y.-L.; Min, J.; Zhou, Y.-Q.; Huang, X.-J.; Wu, L.-D. Semantic orientation computing based on hownet. *J. Chin. Inf. Process.* **2006**, *20*, 14–20.
28. Zhang, C.; Liu, P.; Zhu, Z.; Fang, M. A sentiment analysis method based on a polarity lexion. *J. Shandong Univ. Nat. Sci.* **2012**, *47*, 47–50.
29. Wang, Z.; Yu, Z.; Guo, B.; Lu, X. Sentiment analysis of Chinese microblog based on lexicon and rule set. *Comput. Eng. Appl.* **2015**, *51*, 218–225.
30. Liu, P.Y.; Zhang, Y.H.; Zhu, Z.F.; Xun, J. Micro-blog orientation analysis based on emoticon symbol. *J. Shandong Univ.* **2014**, *49*, 8–13.
31. Weibo. Available online: https://weibo.cn/pub/ (accessed on 5 March 2018).
32. An Open Source and Collaborative Framework for Extracting the Data You Need from Websites. Available online: https://scrapy.org/ (accessed on 3 July 2018).
33. FreeICTCLAS. Available online: https://www.oschina.net/p/freeictclas (accessed on 12 December 2018).