Post-print of Research on Stance Detection in Chinese Microblogs

Authors: Liu Kan, Tian Ningmeng, Wang Hongyu, Lin Rongrong, Wang Demin

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] This study proposes a combined classification model (SWNB model) based on a sentiment weighting algorithm and Naive Bayes algorithm for stance detection on Chinese microblog topics. [Method/Process] The model first simplifies microblogs through a given complex sentence model, then obtains sentiment weights based on sentiment rules, extracts and optimizes topic-related entities from microblogs, thereby classifying them into two categories: those containing a stance and those with no stance indicated (NONE). Subsequently, for microblogs containing a stance, it extracts feature words and utilizes the Naive Bayes algorithm to classify their stance as either support (FAVOR) or opposition (AGAINST). [Results/Conclusion] Experimental results demonstrate that the proposed model achieves favorable accuracy in stance detection and can effectively handle complex scenarios such as Chinese complex sentence patterns, topic-related evaluation objects, and contextual information.

Full Text

Stance Detection in Chinese Microblogs

Liu Kan, Tian Ningmeng, Wang Hongyu, Lin Rongrong, Wang Demin
Department of Information, School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073

Abstract

[**Purpose/Significance**] This paper proposes a combined classification model (SWNB model) based on a sentiment weighted algorithm and Naive Bayes algorithm to detect stance in Chinese microblog topics. [**Method/Process**] The model first simplifies microblogs using a defined complex sentence pattern, then

obtains sentiment weights based on sentiment rules, extracts and optimizes topic-related entities, and classifies microblogs into those containing stance and those with no stance (NONE). For microblogs containing stance, feature words are extracted, and Naive Bayes algorithm is used to classify their stance as support (FAVOR) or opposition (AGAINST). [**Result/Conclusion**] Experimental results show that this model achieves good stance detection accuracy and can effectively handle complex situations such as Chinese complex sentence patterns, topic-related evaluation objects, and contextual semantics.

Keywords: Chinese microblogs, stance detection, sentiment weighted algorithm, Naive Bayes

1 Introduction

In recent years, social media has flourished. Microblog platforms such as Twitter and Sina Weibo have gradually become global centers for disseminating hot information due to their timely and convenient interactive methods, simple and diverse operation modes, and efficient and open communication methods. More and more users choose to express their emotional experiences and comment on current events on microblogs, which contain rich emotional information. Therefore, stance detection has emerged as a research hotspot in the field of text orientation analysis. Stance detection refers to the ability to automatically determine whether the text author holds a supportive (FAVOR), opposing (AGAINST) stance, or does not express any stance (NONE) regarding a given target topic [**?**]. Timely grasping the stance on hot topics in microblogs and effectively extracting user emotional information has become a powerful tool for public opinion analysis, public opinion supervision, and enterprise product experience improvement.

2 Related Work

Stance detection, although part of text orientation analysis research, differs from traditional sentiment analysis. Traditional sentiment analysis analyzes subjective texts with emotional color or evaluative tendency, mines opinions within them, and directly obtains evaluation information about objects contained in the text [**?**]. However, stance detection emphasizes whether the text holds a supportive or opposing stance toward a given target topic. The text itself may not contain any emotional color or subjective evaluation, or may indirectly express the author' s stance on the given target topic by expressing views on other events. Therefore, the given target topic may not explicitly appear in the text [**?**].

Current stance detection research methods proposed by scholars mainly build upon the following two types of text orientation analysis methods:

(1) **Rule-based analysis**. This approach extracts sentiment factors from text using resources such as sentiment dictionaries, considers the depen-

dency relationships between sentiment factors and feature objects, and obtains the overall text sentiment tendency through simple statistics of tendency values. However, such methods often fail to capture implicit textual semantic relationships. Representative works include: Y. Lu et al. proposed an optimization method for automatically constructing context-dependent sentiment dictionaries based on unified principles [**?**]; C. J. Hutto et al. proposed the VADER rule model, which comprehensively considers emotional knowledge, grammatical structure, and semantic features of English text to make fine-grained distinctions in sentiment intensity [**?**]; Chen Yijin et al. proposed a quantitative calculation method for public opinion sentences that can extract the theme of public opinion sentences and perform sentiment tendency analysis on posts regarding that theme [**?**]; Liu Quanchao et al. constructed sentiment analysis dictionaries, internet slang dictionaries, and emoticon libraries based on microblog content features and forwarding/comment relationship features, designing a microblog topic sentiment tendency determination algorithm based on phrase paths [**?**]; B. K. Y. Tsou et al. analyzed the sentiment tendency of news text by calculating the semantic orientation of words and comprehensively considering the distribution, density, and semantic intensity of polarity elements [**?**].

(2) **Machine learning methods**. These methods construct classifiers using machine learning models on the basis of feature extraction, thereby transforming text orientation analysis into a classification problem. However, such methods cannot well consider the influence of sentence patterns and contextual factors. For example, M. Wojatzki et al. used a Stacking-based combined classification method, extracting n-gram, syntactic, lexicon, target-transfer, concept and other features, then adopted a trainable meta-learning method to combine multiple base classifiers to achieve stance detection [**?**]; P. Anand et al. proposed a stance detection model for online real-time discussions, using the JRIP algorithm for continuous rule induction learning, extracting relevant features according to rules, and then classifying using Naive Bayes algorithm [**?**]; S. M. Mohammad et al. extracted semantic and sentiment features from text based on manually constructed word-granularity sentiment dictionaries and sentiment symbol dictionaries to construct feature vectors, using SVM algorithm to detect sentiment tendency of specific evaluation objects in Twitter text [**?**]; B. Velichkov et al. used the GATE framework to extract feature information, then adopted a linear SVM model to classify feature vectors [**?**]; W. Casey et al. defined evaluation information phrases for Twitter data and defined five attributes for them: attitude, orientation, force, focus, polarity, extracting attribute features from text as input for SVM classifier [**?**]; A. Severyn et al. used a convolutional neural network model for sentiment analysis, taking character-level word vectors as raw features, concatenating feature vectors using multiple convolutional kernels of different sizes, achieving high accuracy [**?**].

Regarding stance detection problems, detecting stance in Chinese microblogs is more difficult than in English microblogs, mainly for the following reasons: (1) Word segmentation is a key step in Chinese text analysis, and the quality of segmentation results directly affects model accuracy; (2) Microblog expression is relatively casual, and microblog authors often spontaneously generate related internet slang and nicknames for a certain topic, such as "native chicken," "burn high incense," "cheat people," etc.; (3) Semantic relationships in microblog text are more implicit, and research shows that traditional dependency parsing methods are not suitable for extracting evaluation objects and evaluation words from microblog text [**?**]. Given these problems, this paper proposes a SWNB (serial sentiment weighted and naïve bayes model) model that transforms the three-classification problem into multiple binary classification problems based on combining sentiment weighted algorithm and Naive Bayes algorithm. Using a semi-supervised learning method, it expands the sentiment lexicon and establishes associated entity sets for specific topics to help improve the accuracy of evaluation object extraction. It proposes sentiment weighted rules that can simultaneously process complex sentence patterns and topic-related entities, effectively distinguishing whether the text expresses a stance. The Naive Bayes algorithm focuses on various semantic elements in microblogs such as sentiment words, negation words, conjunction words, degree adverbs, etc., to conduct detailed stance discrimination.

3 Feature Engineering

3.1 Sentiment Feature Words

This paper references the NTUSD simplified Chinese sentiment dictionary from National Taiwan University (http://nlg.csie.ntu.edu.tw/) and the HowNet Chinese sentiment dictionary (http://www.keenage.com/) to construct a sentiment word list, filtering out words with ambiguous sentiment tendencies. Since existing sentiment dictionary resources are not targeted, some verbs and sentiment words only exhibit certain sentiment tendencies when appearing in contexts related to a specific target topic. These words should also be considered sentiment items. For example, in microblogs related to the topic "iPhone SE," verbs such as "buy" and "purchase" often appear, indicating the author' s positive attitude toward the described object; in microblogs related to the topic "Spring Festival firecrackers," adjectives such as "vivid" and "full of festive atmosphere" often appear, indicating the author' s love for this custom. Therefore, this paper manually supplemented some verbs and adjectives with positive or negative sentiment tendencies related to each target topic.

3.2 Associated Entities

Microblog authors directly or indirectly express their stance on a target topic by evaluating entity objects directly or indirectly related to that topic. This paper defines these entity objects as associated entities. Effectively identifying associated entities in text and analyzing the sentiment attitude of microblog authors

toward these entities to determine their stance on the target topic makes stance detection more targeted and improves discrimination accuracy. In constructing the associated entity database, this paper defines the following basic terms and data structures:

(1) **Core Entity**: Represents the core content (people, things, organizations) directly related to the target topic (Target). The core entity set corresponding to a certain target topic can be represented as Target: {core entity 1, core entity 2, core entity 3, ···}.

(2) **Normal Entity**: Represents content indirectly related to the target topic (Target) but in a comparison/parallel relationship with the core entity of that target topic. If two entities have a comparison relationship, they usually exhibit different sentiment tendencies; if they have a parallel relationship, they usually exhibit similar sentiment tendencies. Normal entities corresponding to a certain Target can be represented according to the following structure:

```
Entity Name (Normal Entity Name)
Target Topic (Target)
Corresponding Core Entity (Corresponding core entity): From core entity set
Relationship with Core Entity (Relationship with core entity): Comparison relationship / Par
Sentiment Tendency: Positive (positive) / Negative (negative)
```

The associated entity database corresponding to each target topic contains the core entity set of that target topic and the normal entity set divided according to comparison/parallel relationships. In constructing the associated entity database, this paper uses the keyword recognition function of the NLPIR system to extract keywords from microblogs of each target topic, uses the part-of-speech tagging function of the NLPIR system (http://ictclas.nlpir.org/) to extract nouns from microblogs of each target topic, conducts word frequency statistics on these words, and supplements with manual screening to obtain the core entities corresponding to each target topic. Then, using each core entity as a seed word, for microblogs containing that core entity, according to the four types of comparative sentence patterns proposed by Song Rui et al. [**?**], it identifies microblogs containing comparative relationships, uses the Conditional Random Field (CRF) model [**?**] commonly used in sequence labeling to extract comparative subjects, comparative objects and their parts of speech and positions. If the extracted subject or object contains a core entity, the other non-core entity is defined according to the above data structure and added to the normal entity set. For microblogs not containing comparative relationships, it uses the dependency parsing tool in the Harbin Institute of Technology Language Technology Platform (http://www.ltp-cloud.com/) to label parallel relationship components in microblogs, extracts entities that have a parallel relationship with the core entity, defines them according to the above data structure and adds them to the normal entity set. For some entities that cannot be determined, manual assistance is used for screening. Taking the dataset used in this paper' s experiments as an example, the dataset contains five topics:

"Spring Festival firecrackers," "iPhone SE," "Russia' s anti-terrorism operations in Syria," "Open second child policy," and "Shenzhen motorcycle and electric bike ban." Corresponding associated entity databases were constructed for each target topic, as shown in Table 1 :

3.3 Clause Conjunctions

According to grammatical relationships, complex sentences can usually be divided into transitional, conditional, hypothetical, causal sentences, etc., each with specific clause conjunctions. Conjunctions such as "although," "no matter," "even if" usually lead clauses that are opposite to the author' s true feelings, and are often called "concession conjunctions." Conjunctions such as "but," "however" usually lead clauses that express the same emotion as the author' s true feelings, and are often called "insistence conjunctions" [**?**].

Table 2 Conjunction Table

Type	Common Conjunctions
Concession conjunctions	although, despite, admittedly, no matter, regardless, irrespective, even if, even though, granting, granted, even, even when, would rather, even supposing
Insistence conjunctions	but, however, yet, nevertheless, but, yet, however, not as good as, also

3.4 Intensity Modifiers

The intensity of sentiment words is affected by the modification of adverbs and negation words [**?**]. If a negation word modifies a sentiment word in the text, the sentiment tendency expressed by the text will be reversed. Therefore, this paper collects some commonly used negation words for identifying negative sentences.

If a degree adverb modifies a sentiment word in the text, the intensity of the sentiment expressed by the text differs. This paper divides degree adverbs into six levels: most |most, very |very, more |more, slightly |-ish, insufficiently |insufficiently, over |over, setting the weights corresponding to each level of degree adverbs as 2, 1.25, 1.2, 0.8, 0.5, 1.5 respectively, to make fine distinctions among different intensities of sentiment tendencies.

4 Stance Detection Model

Viewing stance detection as a classification problem, this paper' s SWNB model transforms the three-stance classification into multiple binary classification problems [**?**]. First, a new sentiment weighted algorithm is used to classify microblogs into those containing stance (non-NONE) and those with no stance (NONE). Then, Naive Bayes algorithm is used for binary classification of microblogs classified as containing stance (non-NONE) by the first layer classifier, dividing their

stance into support (FAVOR) or opposition (AGAINST). The overall framework of the SWNB model is shown in Figure 1 [Figure 1: see original paper].

4.1 Sentiment Weighted Algorithm

When calculating the sentiment weight of microblog text, first the microblog text is segmented into sentences according to punctuation marks such as '。' , '!' , '……' , and '?' , transforming the microblog text into a series of sentences. Then, based on the sentence pattern model proposed in this paper, complex sentences are transformed into simple sentences using clause conjunctions. Next, the sentiment weighted algorithm is used to calculate the sentiment weight of each sentence. Then, by detecting whether associated entities related to the target topic appear in the microblog and the relationship between these entities and the core entities of the target topic, the sentiment weight of the sentence obtained in the previous steps is adjusted. Finally, the average of all sentence sentiment values is taken as the sentiment weight of the microblog.

In the classifier training stage, this paper uses a boundary detection method based on grid search algorithm [**?**] to find the optimal upper threshold and lower threshold for dividing non-NONE and NONE sentiment weights (these thresholds maximize the classification accuracy of training data). In the classifier application stage, when the final sentiment weight of the text to be classified falls within the interval formed by the upper and lower thresholds, the stance of the text to be classified is divided into NONE; otherwise, the stance of the text to be classified is non-NONE. The flowchart of the sentiment weighted algorithm is shown in Figure 2 [Figure 2: see original paper].

4.1.1 Complex Sentence Processing Strategy Since clauses led by concession conjunctions are often opposite to the author' s true feelings, and clauses led by insistence conjunctions are often the same as the author' s true feelings, sentiment analysis only needs to be performed on one of them [**?**]. Generally, complex sentence patterns have the following manifestation:

```
[concession conjunction + negation word + sentiment word + punctuation +] insistence conjunc
```

Scanning each sentence Sn in the microblog, first check whether an insistence conjunction appears. If no insistence conjunction appears, directly calculate the sentiment value according to the sentiment rules below. If an insistence conjunction appears, scan the text from the beginning of the sentence to the insistence conjunction. If this part contains a concession conjunction, set the sentiment value of the clause led by the concession conjunction to 0, and calculate the sentiment value of other clauses in this part according to the sentiment rules below. If this part does not contain a concession conjunction, set the sentiment value of the text from the beginning of the sentence to before the insistence conjunction to 0.

4.1.2 Sentence Sentiment Weight Calculation Rules After simplifying complex sentences, the sentiment value of a complete sentence can be directly obtained by calculating the sum of the sentiment values of each clause, while the sentiment value of a clause is obtained based on the sentiment values of each sentiment meaning group in the clause. Each time a sentiment word appears in a clause, it is considered that a sentiment word meaning group appears. Negation words and degree adverbs also affect the intensity of emotion expressed by the meaning group, so negation words and degree adverbs appearing in the context need to be considered when calculating sentence sentiment weights.

(1) Extract sentiment word meaning groups in the clause, representing the relevant information of sentiment words in the following form:

```
senWord = (position in sentence, sentiment tendency, sentiment weight)
```

where the weight of positive sentiment words is set to 1, and the weight of negative sentiment words is set to -1.

(2) Use the position of the previous sentiment word meaning group (lastWordPos) or the position of the previous punctuation mark (lastPuncPos) as the starting point (choose the position closest to the current sentiment word meaning group), and scan between the starting point and the current sentiment word meaning group:

a. Extract degree adverbs, representing the relevant information of degree adverbs in the following form:

```
degreeWord = (position in sentence, weight)
```

b. Extract negation words: When the negation word position precedes the degree adverb position, assign the negation word weight as -1; otherwise, assign the negation word weight as 0.5. If multiple negation words appear in the clause, when the number of negation words is odd, the negation word weight remains unchanged; when the number is even, the negation word weight takes the opposite number.

(3) The sentiment weight of the sentiment word meaning group is calculated using formula (1):

$$\text{scoreW} = \text{negWord} \times \text{degreeWord} \times \text{senWord} \tag{1}$$

The sentiment weight of the clause is the sum of the sentiment tendency values of each sentiment word meaning group in the clause. The sentiment value nsW of a complete sentence nS is obtained by summing the sentiment weights of each clause using formula (2):

$$nsW = \sum \text{subsentence scoreW} \tag{2}$$

After completing the above three steps, the sentiment value of a microblog text is the average of the sentiment values of each sentence.

4.1.3 Associated Entity Processing Strategy Scanning each sentence of the microblog text, if the sentence contains a core entity of the target topic, the sentence sentiment tendency value remains unchanged; if the sentence contains a normal entity of the target topic and the normal entity is defined as having positive sentiment tendency, the sentence sentiment tendency value remains unchanged; if the sentence contains a normal entity of the target topic and the normal entity is defined as having negative sentiment tendency, the sentence sentiment tendency value takes the opposite number; if the sentence contains neither core entity nor normal entity, the sentence sentiment tendency value is not changed.

4.2 Naive Bayes Algorithm

Naive Bayes algorithm is used for binary classification (support or opposition) of microblogs classified as non-NONE by the sentiment weighted algorithm. Extracting sentiment words, associated entities, negation words, and degree adverbs from each microblog text based on the sentiment lexicon, negation word list, degree adverb list, and associated entity database of each target topic, these are used as feature words to calculate the joint probability of these feature items and each category, thereby estimating the classification probability of a given microblog text. This paper adopts the Bernoulli model [**?**] in Naive Bayes classifier to determine the category C to which microblog X belongs.

Selecting n feature words, using maximum likelihood estimation [**?**] to calculate $P(X|C)$, if a feature word has never appeared in the training set, it will cause the overall probability calculation result to be 0. Therefore, Laplace smoothing is used to smooth its probability value by adding one. In addition, the result of multiplying multiple $P(X|C)$ probability values is very small. Whether high calculation accuracy can be guaranteed when probability values are very small will affect the results. Therefore, data transformation is needed to make its presentation better approach the desired assumption and conduct more accurate statistical inference. Based on Naive Bayes algorithm, by taking the logarithm of $P(X|C)$, the multiplication calculation of probability values can be converted into addition calculation, and uncertainty analysis can be converted into information quantity analysis, thereby improving calculation accuracy and classification correctness. The calculation method is shown in formula (5):

$$C = \arg\max \left[\log P(C) + \sum \log P(X_i|C)\right] \tag{5}$$

5 Experiments

5.1 Experimental Data

This paper selects part of the corpus provided in the 2016 NLPCC evaluation task for stance detection as the experimental dataset, including 3,000 annotated training corpora and 1,000 gold standard test corpora (gold data). Both types of corpora contain microblog data on five target topics: "iPhone SE," "Spring Festival firecrackers," "Russia' s anti-terrorism operations in Syria," "Open second child policy," and "Shenzhen motorcycle and electric bike ban." Table 3 and Table 4 respectively count the data distribution of each target topic in the two corpora.

5.2 Experimental Preprocessing

In the data cleaning process, regular expressions are added to remove @ marks, forwarding marks (usually starting with //@), and web link marks (usually starting with http) from microblog content. Since some microblog texts are news content related to each Target, and the 【】 symbol usually contains key content of the news that can indicate the stance of the microblog, for such texts, only the content within the 【】 symbol is extracted for analysis.

In the word segmentation process, the associated entity database, sentiment word list, clause conjunction list, negation word list, and degree adverb list mentioned in Section 3 are integrated and added to the NLPIR word segmentation system of the Chinese Academy of Sciences as a user-defined dictionary to improve word segmentation effectiveness. In the stop word removal process, the stop word list published by the Information Retrieval Center of Harbin Institute of Technology is used to match and query stop words in microblogs and remove them.

5.3 Experimental Design

As a comparison, the following three models are selected to perform stance detection on the same dataset:

(1) **Naive Bayes three-classification model (NB model)**: Using the Bernoulli model of the Naive Bayes classifier proposed above to classify the stance of a microblog into FAVOR, AGAINST, or NONE.

(2) **SVM three-classification model**: Representing each microblog with a feature vector as input to the SVM algorithm to identify three stances [**?**]. Table 5 lists all feature types and meanings of the SVM model.

(3) **Glove_{SVM} model**: Literature [**?**] uses word vectors trained by unsupervised GloVe algorithm and sums them to obtain the vector representation of microblog text, using it as input to a logistic regression model.

When evaluating model experimental effectiveness, this paper uses accuracy, recall, F-score and other indicators to evaluate the classification results of Favor

and Against stances. In implementation, classification work involving SVM algorithm uses the LibSVM toolkit (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) developed by National Taiwan University.

5.4 Results and Analysis

5.4.1 Experiment 1 Two sets of experiments were conducted for each model. Experiment 1 randomly split the data corresponding to the five target topics in the 3,000 annotated training corpora into two parts according to a 6:4 ratio, used as training set and test set respectively. The final training set contained a total of 1,800 data items (360 data items for each target topic), and the test set contained a total of 1,200 data items (240 data items for each target topic).

Table 6 lists the overall F-score of the test data and the F-score corresponding to each target topic in Experiment 1. The line chart in Figure 3 [Figure 3: see original paper] visualizes the overall experimental results of the test data across models, and the bar chart visualizes the experimental results corresponding to each target topic.

From the experimental results, it can be found that except for the topic "Russia' s anti-terrorism operations in Syria," the F-score of the SVM three-classification model is lower than that of the NB model and Glove_{SVM} model, indicating that the quality of feature selection has a great impact on the classification effect of the SVM model. On the one hand, using word bigrams as text features results in large data scale; on the other hand, features such as the number of negation words, the number of degree adverbs, and the number of associated entities cannot capture the sentiment modification relationships among the three. The GLOVE_{SVM} algorithm trains word vectors through GloVe algorithm, obtains text vectors, and considers potential semantic relationships at different granularities in the text through deep learning, enabling better representation of text features.

Naive Bayes algorithm directly classifies the stance of a microblog into FAVOR, AGAINST, or NONE, with a relatively high overall F-score, indicating that the Naive Bayes model has high accuracy in identifying FAVOR and AGAINST stances.

Compared with NB model, SVM model, and Glove_{SVM} model, the SWNB model proposed in this paper achieves more accurate stance detection results for the five target topics. When the SWNB model classifies microblog stances into containing stance (non-NONE) and no stance (NONE), it considers whether entities related to the topic appear in the microblog text, and then analyzes the sentiment attitude of microblog authors toward these entities to determine their stance on the target topic. In microblogs with NONE stance, authors often only conduct objective analysis of the corresponding topic facts without expressing any attitude. For example, the microblog "Open second child policy is both a test of young people' s courage and a test of grandparents' energy. Chinese parents must work hard to raise their own children and also take great pains to

care for their children' s children, not easy." does not reveal the author' s clear stance on the "open second child policy." The improved Naive Bayes algorithm focuses on the influence of sentiment feature words, associated entities, negation words, and degree adverbs when performing binary classification on microblogs classified as containing stance (non-NONE).

5.4.2 Experiment 2 To further verify the effectiveness and rationality of the model proposed in this paper, Experiment 2 uses the 3,000 annotated training corpora as the training set and the 1,000 gold standard test corpora (gold data) as the test set for experiments. Table 7 lists the overall F-score of the test data and the F-score corresponding to each target topic in Experiment 2. The line chart in Figure 4 [Figure 4: see original paper] visualizes the overall F-score of the test data, and the bar chart visualizes the F-score corresponding to each target topic.

The SWNB model proposed in this paper can achieve an F-score of 0.7 when identifying the stance of microblogs on the two target topics "iPhone SE" and "Spring Festival firecrackers" ; the F-score for the topic "Open second child policy" is the lowest at 0.52; the F-scores for the other two target topics are between 0.52-0.7. Analysis of the results found that among microblogs with incorrect stance identification, most microblogs on the topics "Spring Festival firecrackers," "iPhone SE," and "Russia' s anti-terrorism operations in Syria" were identified as having the opposite stance of the correct result, mainly because these microblogs often use irony and other methods to express stance. For example, "Playing dumb? Russian currency fell 60%, oil economy is dying, GDP fell 3.7% in 2015, feeling that it' s all part of Russia' s big chess game?" expresses opposition to Russia' s anti-terrorism operations in Syria through rhetorical questions. Therefore, when only extracting sentiment words, associated entities, negation words, and degree adverbs as feature words in the second layer classifier, such cases cannot be well identified. Most microblogs on the topics "Open second child policy" and "Shenzhen motorcycle and electric bike ban" were classified as NONE, mainly because these topics' microblogs directly contain associated entities at low frequency, leading to errors in sentiment weight calculation in the first layer classifier.

6 Conclusion

This paper proposes a supervised SWNB classification model to detect stance in Chinese microblog topics. The SWNB model proposes new sentiment weighted rules for processing complex sentence patterns and topic-related entities, which can effectively distinguish whether the text expresses a stance. The improved Naive Bayes algorithm can perform binary classification on microblogs classified as containing stance (non-NONE) by the sentiment weighted algorithm, dividing them into support (FAVOR) or opposition (AGAINST) stances. The SWNB model proposed in this paper combines the advantages of sentiment rules and machine learning models, fully considering the influence of Chinese complex sen-

tence patterns, topic-related entities, contextual semantics, and text semantics on text sentiment tendency, with simple implementation and high detection accuracy.

However, the SWNB model heavily depends on the completeness of resources such as sentiment dictionaries and associated entity sets, as well as the accuracy of word segmentation results. In addition, a large number of sentiment words are ambiguous, and their expressed meanings differ in different contexts. The current model cannot strictly distinguish ambiguous sentiment words. Therefore, future work needs to further improve resources such as sentiment dictionaries and combine semantic analysis and deep learning techniques to more accurately detect stance in Chinese microblog topics.

References

[1] MOHAMMAD S M, SOBHANI P, KIRITCHENKO S. Stance and sentiment in Tweets[J]. ACM transactions on embedded computing systems, 2016, 1(1):1-21.

[2] MANKE S N, SHIVALE N. A review on: opinion mining and sentiment analysis based on natural language processing[J]. International journal of computer applications, 2015, 109(4): 29-32.

[3] BøHLER H, ASLA P F, MARSI E, et al. IDI@NTNU at SemEval-2016 task 6: detecting stance in Tweets using shallow features and GloVe Vectors for word representation[C]//International workshop on semantic evaluation. San Diego: Elsevier, 2016: 445-450.

[4] LU Y, CASTELLANOS M, DAVAL U, et al. Automatic construction of a context-aware sentiment lexicon: an optimization approach[C]//International conference on World Wide Web. Hyderabad: ACM, 2011: 347-356.

[5] HUTTO C J, GILBERT E. VADER: A parsimonious rule-based model for sentiment analysis of social media text[C]// Eighth international conference on weblogs and social media. Michigan: AAAI, 2014: 216-225.

[6] Chen Yijin, Cao Shujin, Chen Guihong. Opinion mining for online public opinion: research on sentiment orientation analysis of user comments[J]. Knowledge of Library and Information Science, 2013(6): 90-97.

[7] Liu Quanchao, Huang Heyan, Feng Chong. Research on sentiment orientation determination algorithm for microblog topics based on multi-features[J]. Journal of Chinese Information Processing, 2014, 28(4): 123-131.

[8] TSOU B K Y, YUEN R W M, KWONG O Y, et al. Polarity classification of celebrity coverage in the Chinese press[C]// 2005 International Conference on Intelligence Analysis. Virginia: IEEE, 2005: 102-108.

[9] WOJATZKI M, ZESCH T. ltl.uni-due at SemEval-2016 Task 6: Stance detection in social media using stacked classifiers[C]// International workshop on

semantic evaluation. San Diego: Elsevier, 2016: 428-433.

[10] ANAND P, WALKER M, ABBOTT R, et al. Cats rule and dogs drool!: Classifying stance in online debate[C]// The 2nd workshop on computational approaches to subjectivity and sentiment analysis. Oregon: ACL Anthology, 2011: 1-9.

[11] MOHAMMAD S M, KIRITCHENKO S, Zhu X. NRC-Canada: building the state-of-the-art in sentiment analysis of tweets[C]//International workshop on semantic evaluation. Atlanta: ACL Anthology, 2013: 321-327.

[12] VELICHKOV B, KAPUKARANOV B, GROZEV I, et al. SU-FMI: System description for SemEval-2014 Task 9 on sentiment analysis in Twitter[C] // International workshop on semantic evaluation. Dublin: ACL Anthology, 2014: 73-78.

[13] CASEY W, NAVENDU G, SHLOMO A. Using appraisal groups for sentiment analysis[C]// The 14th ACM international conference on Information and knowledge management. Bremen: ACM, 2005: 625-631.

[14] SEVERYN A, MOSCHITTI A. UNITN: Training deep convolutional neural network for Twitter sentiment classification[C]// The 9th international workshop on semantic evaluation. Denver: ACL Anthology, 2015: 464-469.

[15] LIU X, LI K, ZHOU M, et al. Collective semantic role labeling for tweets with clustering[C]// The twenty-second international joint conference on artificial intelligence. Barcelona: AAAI, 2011:1832-1837.

[16] Song Rui, Lin Hongfei, Chang Fuyang. Recognition of Chinese comparative sentences and extraction of comparative relations[J]. Journal of Chinese Information Processing, 2009, 23(2): 102-107.

[17] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C] //Proceedings of ICML 2001. San Francisco, 2001: 282-289.

[18] Yao Shuangyun, Shen Wei. Research on collocation of conjunctions[J]. Computer and Modernization, 2007(4): 7-9.

[19] Liu Cuijuan, Liu Zhen, Chai Yanjie, et al. Research on visual computing method for social group emotion based on microblog text data analysis[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 178-186.

[20] Han Hong, Yang Jingyu, Lou Zhen. Hierarchical classifier combination[J]. Journal of Nanjing University of Science and Technology (Natural Science), 2002, 26(1): 10-14.

[21] LAMESKI P, ZDRAVEVSKI E, MINGOV R, et al. Svm parameter tuning with grid search and its impact on reduction of model over-fitting[M]//In rough sets, fuzzy sets, data mining, and granular computing. Germany: Springer, 2015: 464-474.

[22] Li Aiping, Di Peng, Duan Liguo. Document sentiment analysis based on sentence sentiment weighted algorithm[J]. Journal of Chinese Computer Systems, 2015, 36(10): 2252-2256.

[23] LEWIS D D. Naive (Bayes) at forty: the independence assumption in information retrieval[C]//European Conference on Machine Learning. Berlin: Springer, 1998: 4-15.

[24] M Y U N G I J. Tutorial on maximum likelihood estimation[J]. Journal of mathematical psychology, 2003, 47(1): 90-100.

[25] PANG B, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques[C]// The ACL-02 conference on empirical methods in natural language processing. Philadelphia: ACL Anthology, 2002: 79-86.

Author Contributions

Liu Kan: Mainly completed the research idea, algorithm model proposal, and paper revision.
Tian Ningmeng: Mainly completed the algorithm proposal, main experiments, and initial draft writing.
Wang Hongyu: Mainly completed algorithm discussion and comparative experiments.
Lin Rongrong: Mainly completed comparative experiments.
Wang Demin: Mainly completed data collection and preprocessing.

Note: Figure translations are in progress. See original paper for figures.