Emotion recognition method of Tibetan micro-blog text based on sentiment dictionary

Benwang Sun\textsuperscript{1}, Fang Tian\textsuperscript{1}, Mengmeng Jia\textsuperscript{1}

\textsuperscript{1}Department of Computer Technology and Applications, Qinghai University, Xining, China

\textsuperscript{*}Fang Tian’s e-mail: tianfang@qhu.edu

Abstract: With the development of network technology, social media represented by micro-blog has become an important platform for personal emotional expression. Automatic analysis of micro-blog texts and identification of their emotions is important for understanding people’s emotions. In this paper, the emotion of Tibetan micro-blog text is analyzed automatically for the first time, the emotion in text expression is recognized, the analysis method based on Tibetan sentiment dictionary and rules is put forward, and the language characteristics of emotion expression in Tibetan micro-blog are discussed. The experimental results show that the accuracy of emotion recognition based on Tibetan sentiment dictionary reaches 78.6\%, which provides a basis for establishing a high-precision Tibetan textual sentiment classification system.

1. Introduction

With the rapid development of the Internet, online social media has become the main carrier of personal information dissemination in people's daily life, among which micro-blog platform is a typical representative of many online social applications. Emotion recognition for micro-blog text is a basic research in the field of natural language processing, and has received close attention from scholars in recent years. Micro-blog text emotion recognition is to analyze micro-blog texts based on sentiment dictionary or machine learning, and to identify whether the text contains emotional expression, which is also the basis and premise of sentiment classification. In massive micro-blog texts, there are a large number of emotional texts and non-emotional texts. Among them, the non-emotional text does not carry any people's emotional color. Text content is only an objective description of individuals and things, which will have an interference effect on sentiment classification, while the emotional text contains personal emotions such as sadness, happiness, and miss.

2. Related research

2.1 Research Status of Sentiment Dictionary Construction

Human emotions are rich and changeable, for the study of sentiment classification, different division systems are quite different. In the sub-category of English micro-blog, Ekman proposed six basic emotional states by studying human facial expressions: joy, anger, fear, disgust and surprise\cite{1}. This classification system is still widely used in the related research of English text sentiment classification. On the basis of Ekman, for the classification of emotion categories in Chinese text, the emotion category "good" was added to the Chinese emotional vocabulary ontology, and the commendatory emotion was divided into more detailed, and finally the emotions were divided into 7 categories and 21...
subcategories in the final vocabulary ontology[2]. The Chinese benchmark sentiment dictionary constructed by Jiao Tong further collected and analyzed the emoticons and network words commonly used in micro-blog text expression, and perfected Micro-blog sentiment dictionary more deeply[3].

2.2 Research Status of Emotion Recognition
The research on emotion recognition had been carried out earlier abroad, and the results were also quite rich. In the early stage, Aman collected blog articles and began the annotation work of word-level and phrase-level emotion categories and intensity, and used knowledge-based methods to identify sentences with or without emotions, and the accuracy rate (Accuracy) reached 73.89 % [4]. Maeda first defined three writing styles, one formal expression and two informal expressions, and conducted emotional recognition research on the Twitter corpus[5]. Esmin used hierarchical classification method on Twitter corpus to automatically analyze the emotional polarity and emotional categories of Twitter texts. Among them, the first layer was the classification of whether there was emotion, that is, trying to determine whether a Twitter text contained emotion, and the latter two layers were the study of emotional polarity and emotional category classification based on the first layer[6]. Jenke used machine learning technology to select features from emotional data sets and compared them experimentally. It was found that the features selected by multivariate methods were superior to univariate feature selection methods in emotion recognition tasks[7].

With the construction and release of Chinese Emotional Corpus, there were more and more researches on emotion recognition in China[8-9]. Yao Yuanlin introduced the relevant results of emotion recognition in NLP&CC2013 Chinese Micro-blog sentiment analysis evaluation task[10]. Liu Huanhuan studied the sentence-level emotion recognition method in the Chinese Emotional Corpus (Ren-CECps) and compared the influence of different features on emotion recognition[11]. Huang Lei proposed a micro-blog text emotion recognition method based on syntactic information, which fully considered the syntactic information of micro-blog text and can obviously improve the performance of micro-blog emotion recognition[12]. Xia Rui proposed a multi-task learning framework based on Deep Belief Network (DBN), which illustrated the benefits of using additional information in multi-task learning of emotion recognition[13]. Yin Hao proposed a micro-blog emotion recognition method based on character and word features, which used Long Short-Term Memory neural network (LSTM) to learn hidden layer features from word feature representation and word feature representation of text respectively. Through the merger of two groups of hidden layer features, the character and word features are obtained, and the emotion recognition was carried out[14].

3. Construction of Tibetan Sentiment Dictionary
Studying the construction of Tibetan sentiment dictionary will be beneficial to the sentiment classification of micro-blogs or other texts, and promote the application research of Tibetan network public opinion analysis and robot emotion recognition.

3.1 Basic information of existing sentiment dictionary
Hownet Dictionary [15] (including positive and negative emotional polarity, 4766 commendatory words, 4370 derogatory words, degree words are divided into the most (most), very (very), more (more), slightly (ish), less (insufficiently), super (over), and negative vocabulary), And Harbin University of Technology’s dictionary of stop words (including 767 stop words). The form of expression of all the words in the Tibetan-Chinese Dictionary [16] are as follows: "XXX motivation launch ideas thoughts..." We collected 27402 pieces of pre-processed data from all the words in the Tibetan-Chinese Dictionary.

3.2 Construction of Tibetan Sentiment Dictionary
The Tibetan sentiment dictionary is constructed based on Hownet Dictionary and Tibetan-Chinese Dictionary. This dictionary contains positive emotion words, negative emotion words, neutral emotion words, degree words, negative words, turning words and so on. The algorithm steps are as follows:
Step 1: Mainly using matching algorithm, using Hownet Dictionary and the Tibetan-Chinese Dictionary to match keywords, and the Sino-Tibetan bilingual sentiment dictionary is obtained.

Step 2: Checking whether the corresponding Sino-Tibetan emotional vocabularies are correct and correct the mistakes in time by the combination of manual and machine proofreading.

Step 3: Extracting the Tibetan emotional vocabularies by using the character encoding range of Tibetan language to obtain the sentiment dictionary.

4. Research methods of emotion recognition

Because the Tibetan corpus is less and not open, this paper uses the artificially segmented Tibetan micro-blog text corpus to carry out the experiment. In the emotion recognition of micro-blog text, firstly, the Tibetan micro-blog text is segmented, then the initial weights of positive and negative emotional vocabularies are assigned to 1, and the neutral emotional vocabularies are assigned to 0; then the degree vocabularies level of the corresponding emotional vocabularies is found, and the degree vocabularies of different levels will give different degree levels. If the micro-blog text contains an odd number of negative vocabularies, the emotion is reversed, and if there is a double negative vocabulary, the emotion should be appropriately enhanced. The micro-blog score finally is the difference of the emotional weight of the whole text. If the difference is greater than zero, micro-blog expresses positive emotion, if the weight difference is less than zero, micro-blog expresses negative emotion, otherwise micro-blog expresses neutral emotion. The algorithm steps are as follows:

Input:
Set micro-blog positive score: positive
Set micro-blog negative score: negative
Set micro-blog positive score: neutral
Set degree word Wi: = {most, very, more, -ish, insufficiently, over};
Set degree level W_Degree(Wa) : = {2, 1.75, 1.25, 0.75, 0.5, 1.5};
Set micro-blog number M : = {M1,M2,M3..., Mk};

Output:
Micro-blog scores: Scores(k);
for each Mk do
for word in Mk do
If (word in stop_word) Delete word;
else if (word in turn_word) Mk = word.next;
else if (word in pos_word) positive=W_Emotion_REsuit(i);
If (word.pre==Wi) positive=W_Emotion_REsuit(i)*W_Degree(Wa);
else if (word in neg_word) negative=W_Emotion_REsuit(j);
If (word.pre==Wi) negative=W_Emotion_REsuit(j)*W_Degree(Wa);
else if (word in neu_word) neutral=W_Emotion_REsuit(q);
If (word.pre==Wi) neutral=W_Emotion_REsuit(q)*W_Degree(Wa);
else if (word.pre==deny_word) positive=(-1)*positive;
negative=(-1)*negative;
neutral=(0)*neutral
else if (word.pre==double-deny_word)
positive=2*positive;
negative=2*negative;
for each Scores(k) do
    Scores=positive-negative;
end
end
end

A rule-based method is proposed to automatically analyze micro-blog text emotion by using dictionary, the purpose of this paper is to study the current situation of Tibetan sentiment dictionary in social media text analysis. In order to reflect the contribution of the dictionary more directly, simple and intuitive rules are established to judge the emotion expressed by a micro-blog. Finally, the algorithm obtains the text of the emotional micro-blog that has been identified, including emotional texts and non-emotion texts.

5. Experimental results and analysis

5.1 Data acquisition
At present, the sentiment analysis of Tibetan micro-blog is still in its infancy, and there is no micro-blog dataset that has already been labeled with emotion categories has been published. To this end, we first manually labeled Sina micro-blog, and established the Tibetan micro-blog emotional annotation data set. We selected some micro-blog texts published daily by ethnic minorities from Sina micro-blog. These micro-blog texts after the crawl were marked by two labeling personnel. The text with the same content as the advertisement and the forwarded content was deleted when the micro-blog text was labeled. Each micro-blog label was divided into emotion and non-emotion.

5.2 Results of the construction of sentiment dictionary
The sentiment dictionary constructed in this paper obtains 5674 commendatory emotional words, 4897 derogatory emotional words, 1437 neutral words, 192 degree words, 17 negative words, 11 turning words and 13 double negative words. The construction of Tibetan sentiment dictionary provides the necessary basic conditions for the emotion recognition of Tibetan micro-blog text.

5.3 Emotional recognition results and analysis
This paper uses 600 Tibetan micro-blog corpora to calculate the micro-blog text emotion through the Tibetan sentiment dictionary, as shown in Table 5.1:

|                     | Correct Number | Error Number |
|---------------------|----------------|-------------|
| Emotional micro-blog| 314            | 86          |
| Non-emotional micro-blog | 158        | 42          |

It can be seen from the table that the accuracy of emotion recognition based on Tibetan sentiment dictionary is still quite high, and the accuracy of non-emotional micro-blog recognition is 0.5% higher than that of emotional micro-blog. In order to improve the accuracy, the characteristics of dictionary and Tibetan language structure should be considered comprehensively.

The comparison of the results of each sentiment dictionary experiment is shown in Table 5.2:

| Dictionary           | Data set                  | Accuracy |
|----------------------|---------------------------|----------|
| Basic emotional      | NLP&CC                    | 53.16%   |
| dictionary [3]       |                           |          |
| Basic emotional      | Popular micro-blog data set | 54.03%   |
| dictionary [3]       |                           |          |
| C-LIWC dictionary [14] | Manually labeled micro-blog | 70.2%   |
It can be seen that the emotion recognition based on the sentiment dictionary mainly depends on dictionary. The quality of the dictionary directly affects the experimental results. Compared with other sentiment dictionaries, the Tibetan sentiment dictionary has more accurate emotional recognition results, which is accurate to 78.6%. Emotion recognition is the basis of sentiment classification. This experiment explores how to improve the accuracy of Tibetan text emotion recognition by means of Tibetan micro-blog and Tibetan sentiment dictionary, combined with the characteristics of Tibetan language.

6. Summary
The Tibetan sentiment dictionary constructed in this paper is more systematic and comprehensive, with a total of 12241 words, including basic emotion words, stop words, negative words, turning words and so on, which will be constantly updated and improved in the future work. The experimental results show that the accuracy of Tibetan micro-blog text emotion recognition based on Tibetan sentiment dictionary reaches 78.6%, and it can be said that Tibetan micro-blog emotion recognition based on dictionary has achieved good results. At the same time, it is proved that the Tibetan sentiment dictionary has certain practical and referential value. There are still some problems, such as the automatic expansion of Tibetan sentiment dictionary and so on.

Acknowledgment
The work is supported by the foundational application research of Qinghai Province Science and Technology Fund (No. 2016-ZJ-743).

References
[1] Ekman P. Facial expression and emotion. American Psychologist, 1993, 48(4):384-392.
[2] Chen Jianmei. The Construction and Application of Chinese Emotion Word Ontology[D]. Dalian University of technology, 2009.
[3] Tong Jiao. Design and implementation of emotional content classification system for micro-blog text[D]. Beijing University of Posts and Telecommunications, 2018.
[4] Aman S, Szpakowicz S. Identifying expressions of emotion in text[A]. Vaclav M, Pavel M. International Conference on Text, Speech and Dialogue[C]. Heidelberg, Berlin: Springer, 2007: 196-205.
[5] Maeda H, Shimada K, Endo T. Twitter Sentiment Analysis Based on Writing Style[M]. Heidelberg, Berlin: Springer, 2012.
[6] Esmin A A A, De Oliveira R L, Matwin S. Hierarchical classification approach to emotion recognition in twitter[A]. Jiawei Han. Machine Learning and Applications (ICMLA)[C]. Florida, USA: IEEE, 2012, 2: 381-385.
[7] Jenke R, Peer A, Buss M. Feature Extraction and Selection for Emotion Recognition from EEG[J]. IEEE Transactions on Affective Computing, 2017, 5(3):327-339.
[8] Xu Linhong, Lin Hongfei, Zhao Jing. Construction and Analysis of Emotional Corpus[J]. JOURNAL OF CHINESE INFORMATION PROCESSING, 2008, 22(1): 116-122.
[9] Song Hongyan, Liu Jun, Yao Tianfang. Construction of an Annotated Corpus for Chinese Opinioned-Subjective Texts[J]. JOURNAL OF CHINESE INFORMATION PROCESSING, 2009, 23(2): 123-128.
[10] Yao Yuanlin, Wang Shuwei, Xu Ruifeng. The Construction of an Emotion Annotated Corpus on Microblog Text[J]. JOURNAL OF CHINESE INFORMATION PROCESSING, 2014,
28(5): 83-91.

[11] Huang Lei, Li Shoushan, Zhou Guodong. Emotion Recognition of Chinese Microblogs with Syntactic Information[J]. COMPUTER SCIENCE, 2017, 44(2): 244-249.

[12] Xia R, Liu Y. A multi-task learning framework for emotion recognition using 2D continuous space[J]. IEEE Transactions on Affective Computing, 2017, 8(1): 3-14.

[13] Hao Yin. Emotion recognition and classification research based on Micro-blog [D]. Soochow University, 2018.

[14] Niu Yun, Pan Minghui, Wei Ou, Cai Xinye. Emotion analysis of Chinese microblog using lexicon-based approach[J]. Computer Science, 2014, 41(9): 253-258, 289.