International Journal of Database Theory and Application
Vol.9, No.9 (2016), pp.149-156
http://dx.doi.org/10.14257/ijdta.2016.9.9.14

A Sentiment Calculation Method Based on Tibetan Semantic Relations

Zhen Zhang and Lirong Qiu

Department of Information Technology, Minzu University of China, Beijing 100081
E-mail: qiu_lirong@126.com

Abstract

Sentiment analysis is shown significant and indispensable status in hot topic, public-opinion poll, knowledge acquiring and recommended goods fields, which is the fundamental work for natural language processing. This paper proposes an approach to build the Tibetan sentiment dictionary and to calculate the sentiment value base on the Tibetan semantic relations. We test our approach on experimental corpus crawled from Sina weibo and the experimental results demonstrate good performance on Tibetan language.

Keywords*: Tibetan Language; Sentiment analysis; Sentiment dictionary*

1. Introduction

The rapid development and widespread use of the Internet has a deep influence on people's life-style, and the number of Internet users has risen sharply in recent two years, and Internet users enjoy the freedom of speech on the network. As a result, the text information content on Internet is getting larger, text form is also becoming diversified, such as product reviews, city news, blog and online forum posts. These unstructured texts contain a lot of valuable information for social stability, e-commerce, market prediction and many more. Therefore, the urgent demand of mining and expressing the deeper semantic information from the mass of unstructured data is showing more significance. Under this background, the research of text emotional tendency has become a hot spot, and has achieved great development. now in the era of pursuing peace, the understanding of the Tibetan people's living habits and ideas have to be read from the Tibetan language; No matter in Tibetan newspapers, or Tibetan related websites, blogs or online reviews, the Tibetan language is used increasingly widely, so it is significant to understand emotional information expressed in Tibetan, not only for the progress of the Tibetan nationality but also for the harmonious development of Sino Tibetan relations. To understand the affective information conveyed in Tibetan sentences, we must firstly understand the semantic and usage of sentiment words in Tibetan, and then through the understanding of emotional words to analyze the feelings in explicit emotional Tibetan sentences, this research is of practical significance.

2. Related Work

The research work on sentiment analysis system began in the early 20th century, PangBo [1] applied the supervised learning based method to sentiment classification of film comment text for the first time in 2002. In the same year, Turney [2] proposed unsupervised sentiment classification method based on semantic tendency. On the basis of the above, two kinds of sentiment analysis methods are derived, which are sentiment analysis method based on supervised learning and sentiment analysis method based on

ISSN: 2005-4270 IJDTA

unsupervised learning.

2.1. The Supervised Learning Method

Sentiment analysis method based on supervised learning is also called method based on machine learning. At present, this method is still the mainstream, in addition to the sentiment analysis based on Non-negative Matrix Tri-factorization [3] and the sentiment analysis based on Genetic Algorithm [4], the most widely used machine learning algorithm is the perceptron, naive Bayes and k nearest neighbor (KNN). The improvement of this kind of sentiment analysis algorithm is mainly focused on the stage of feature selection.

For feature selection, in addition to the n-gram Grammar (n-gram) and part of speech (POS) feature proposed by PangBo, Wilson [5] proposed kinds of syntactic features like hybrid word feature, negative word feature, emotion modification feature and emotion transfer feature. In addition, research in following aspects was also carried out to analyze the sentiment of supervised learning: Melville [6] proposed a method to judge the sentiment polarity of text combined transcendental emotion tends based on sentiment dictionary and sentiment orientation of c posterior training text based on the context. Taboada [7] proposed determine the emotional tendency of the text with the combination of the subject and the feature of the text itself.

At this stage, the supervised learning emotional analysis is relatively mature. The main research ideas drew on the machine learning method such as text classification, but it has not formed a set of independent research method according to its own features. Plus, the number of test sets in the real world is far more than the number of training sets, and the field of test set is not restricted to consistent with the training set as supervised learning, so it has caused some difficulty to this method.

2.2. The Unsupervised Learning Method

Sentiment analysis method based on unsupervised learning can be divided into dictionary based analysis method and rule based analysis method. In addition to Turney (2002), Zhu Yanlan [8] used HowNet to calculate the emotional tendency of Chinese words semantic meaning. Lou Decheng [9] used syntactic structure and dependence relationship to analyze the emotion of Chinese sentence. Hiroshi [10] realize Japanese phrase-level sentiment analysis through a rule-based machine translator. On the basis of SO-PMI algorithm proposed by Turney, Zagibalov [11] analyze the features of Chinese texts and introduce the iterative mechanism, realize the improvement of the accuracy sentiment analysis based on unsupervised learning to a large extent.

1) Dictionary Method

The method based on dictionary makes use of attribute dictionary, emotion dictionary, degree dictionary and negative dictionary to analyze the sentiment orientation of the sentence. The specific method is constructing sentiment lexicon at first, where every sentiment word is given a sentiment polarity and emotional value, and conduct weighted calculation according to emotional words appeared in the sentence and polarity of every sentiment word, thus gain the sentiment score of the sentence. According to this principle, a fuzzy set analysis model is proposed, and the fuzzy set membership function is used to measure the distance between the emotional tendency of the sentence and different levels (positive, negative, neutral).

2) Rule Method

On the basis of using emotion dictionary, the rule-based approach introduce the logical relationships between the components in the statements. It is a method of optimizing the dictionary. The dictionary approach regards every emotion words as a separate unit, estimates the emotional polarity of the sentence according to the polarity and value of sentiment word in the whole sentence, it ignores the logical relationship between the

elements of the sentence. Rule-based approach is to take full account of the relationship between grammatical statements, the sentence is regard as a whole of grammatical structure, rather than the stacking of words.

The method based on unsupervised learning requires a lot of the related linguistic knowledge such as semantic knowledge, and it is very difficult to achieve. In practical applications, it is the semantic structure rule and semantic knowledge that distinguish the sentiment analysis and the text classification. The dictionary method considers the sentiment word as an independent unit, ignores the logical relationships between sentiment words, leads to inaccurate results. Therefore, we proposes an approach based on syntax structure to calculate sentence sentiment value.

3. Tibetan Sentiment Calculation

3.1. The Tibetan Sentiment Dictionary

Currently, Tibetan corpus and dictionary resources are in urgent need, building Tibetan sentiment dictionary cannot directly take the method of building English and Chinese sentiment dictionary. The work of building sentiment dictionary needs to use the third-party dictionaries. SentiWordNet [12-13], it is the English sentiment dictionary that widely recognized in the field of English sentiment analysis, it has been upgraded to version 3.0. Here we combine English sentiment dictionary SentiWordNet with English-Tibetan dictionary to build Tibetan basic sentiment dictionary, and the specific flow is shown in Figure1.

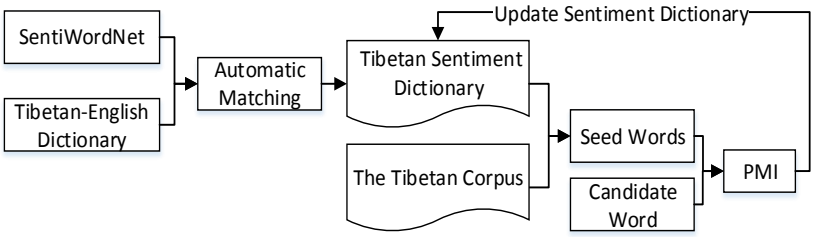


Figure 1. The Construction of the Sentiment Dictionary

That the sentiment words organized in SentiWordNet was showed as Figure2. Because of many meanings of a word, we can't use the value directly, for example, the word "nice" has five meanings as an adjective, one meaning as noun. The calculation formula is shown as the following:

$$\text{Score(word)} = \sum \frac{\quad}{\quad} \tag{1}$$

$$\text{Sum(word)} = \sum \frac{\quad}{\quad} \tag{2}$$

$$\text{S(word)} = Score(word)/Sum(word) \tag{3}$$

Calculate the average as the sentiment value in one part of speech, and the sentiment value maintains at [-1,1], if the value is a positive number, it represents positive polarity，and if the value is a negative number, it represents negative polarity, and if the value is equal to 0.0, it represents neutral polarity.

Figure 2. The SentiWordNet

The Basic Sentiment dictionary is constructed by automatic matching. So, we can get the sentiment value of common sentiment words from the Basic Sentiment dictionary. But as for the unusual sentiment words that called candidate words, we propose an extension method that use the Point-wise Mutual Information (PMI)[14] combined sentiment seed words with the corpus. PMI is a useful method of computational linguistics model analysis which is used to measure the mutuality between two objects. The formula is defined as the following:

$$\mathrm{PMI}(w_1,w_2) = \log_2 \frac{p(w_1 \& w_2)}{p(w_1)p(w_2)} \quad (4)$$

Where w_1, w_2 are two independent words, $p(w_1 \& w_2)$ is the concurrence probability of w_1 and w_2, $p(w_i)$ is the occurrence probability of the w_i. The polarity was calculated by equation (5):

$$\mathrm{SO-PMI}(w) = \sum_{pw \in Pset} \mathrm{PMI}(w,pw) - \sum_{nw \in Nset} \mathrm{PMI}(w,nw) \quad (5)$$

Where $Pset$ and $Nset$ stand for the positive and negative seed words set respectively, w is the candidate word, pw is positive word from the positive seed set $Pset$, nw is negative word from the negative seed set $Nset$.

Here, the experimental procedure is as following:

Step1: Selecting a certain amount of sentiment words with strong polarity from The Basic Sentiment dictionary as part of the seed set;

Step2: Processing corpus with participle and speech tagging, and take a statistics on the word frequency, then choice the higher frequency words with the emotional tendency as another part of the seed set.

Step3: Calculate the sentiment value of the candidate word with the equation (5).

In this section, we construct a Tibetan sentiment dictionary based on the SentiWordNet dictionary, and proposed a method to expand the sentiment dictionary.

3.2. The Sentiment Calculation Model

According to analyzing Tibetan grammar, we divide the sentence into three-layer model according to modification and logical relations between sentences.

The first-layer structure is the simplest modification structure, it is the modification to the sentence backbone with adverbs, adjectives and other words. Noun phrases and verb phrases are the mainly objects.

1) The subject and object are generally consist of the “adjective + noun” structure, both adjectives and nouns have their own sentiment tendencies, the algorithm rules are shown in Table1.

Table 1. Calculation Rules of the Noun Phrases

	adj	n	calculation formula
1.1	+/0	+/0	+(|N|+(1-|N|)*|Adj|)
1.2	+/0	-	-(|N|+(1-|N|)*|Adj|)
1.3	-	+/0	A
1.4	-	-	-(|N|+(1-|N|)*|Adj|)

2）The verb phrase contains a verb and an adverb, that is to say, it consists of the form "adverb verb", adverbs describe the verbs, the calculation rule is shown in Table2.

Table 2. Calculation Rules of the Verb Phrases

	adv	verb	calculation formula
2.1	+/0	+/0	+(|V|+(1-|V|)*|Adv|)
2.2	+/0	-	-(|V|+(1-|V|)*|Adv|)
2.3	-	+/0	Adv
2.4	-	-	-(|V|+(1-|V|)*|Adv|)
2.5	-	+	-(|Adv|*(1-|V|))

The second layer is predicate structure which is mainly consisted of the verb-object relationship. The verb-object relationship is formed by a verb and an object which the verb described. Object structures, which depends more on the action, rely less than verbs. According to this feature, when determining verb-object relationship, we should pay more attention to identifying the verbs. The predicate portion includes a verb phrase, an object or complement. The main form is "verb + object or complement". The calculation rules are shown in Table3.

Table 3. Calculation Rules of the Predicate Relations

	verb	obj/c	calculation formula
3.1	+/0	+/0	+(|O|+(1-|O|)*|V|)
3.2	+/0	-	-(|O|+(1-|O|)*|V|)
3.3	-	+/0	V
3.4	-	-	-(|O|+(1-|O|)*|V|)

The third layer is the relationship between the clauses. Clause is the smallest unit to express a complete semantic meaning, there is a subject-predicate structure in each clause. A subject-predicate structure typically contains subject and predicate. Subjects are generally acted by nouns, noun phrases and pronouns. Verb, adjective, verb phrases, adjectives can be used as predicate. Predicate part is usually expressed as "verb phrase + object / complement". A clause contains a subject and a predicate, the calculation section of subject and predicate are shown in Table3 and Table 3-3, the calculation rules of clause are based on the sentiment score of subject and predicate, the rules is shown in Table4.

Table 4. Calculation Rules of the Clause Relations

	s	v	calculation formula
4.1	+/0	+/0	+(|P|+(1-|P|)*|S|)
4.2	+/0	-	P
4.3	-	+/0	|S|>|P|?|S|:|P|
4.4	-	-	+(|P|+(1-|P|)*|S|)

In addition to the four rules mentioned above, there are some other rules, for these rules we use other calculation rules, as shown in Table5.

Table 5. The Other Calculation Rules

	A	B	calculation formula
5.1	+/0	+/0	+(|A|+(1-|A|)*|B|)
5.2	+/0	-	|A|>|B|?|A|:|B|
5.3	-	+/0	|A|>|B|?|A|:|B|
5.4	-	-	-(|A|+(1-|A|)*|B|)

This calculation order is in line with people's understanding of sentence structure, and closer to sentence processing order of human brain.

4. Experimental Results and Analysis

Experimental procedure is divided into two parts. The first part is the sentence structure analysis that we get the sentence component dependencies by the parsing; the second part is calculation part of the sentiment sentence by the proposed computing model.

4.1. Experimental Data

The paper uses the Sina Weibo Interface to crawl the weibo corpus, a total of 22710 weibo which were deduplicated and removed link tags and useless information. Then choice the 8000 Tibetan weibo as our experimental data that retained the text, emoticons and the related Chinese notes. The data storage format is shown as Table6.

Table 6. The Storage Format of Tibetan Weibo

content	notes
འདི་ཕྱི་གཉིས་ཀྱི་ལམ་སྟོན་ཟ་མ་ལགས་ཁྱེད་རྣམས་ཅི་ཞེ་བསམ་བཀྲ་བདུན་ལོ།	[祈祷][祈祷][祈祷]
ཨོཾ་ཨཱཿཧཱུྃ་བཛྲ་གུ་རུ་པདྨ་སིདྡྷི་ཧཱུྃ། ཟ་མ་མཁྱེན་ནོ།	[祈祷]
ཤས་ཤས་ཅི་མི་སྐྱོན་ ཁྱེད་ཀྱི་བསྟགས་བརྗོད་ལ་ཐུགས་རྗེ་ཆེ་	[可爱] [太开心]
རིན་པོ་ཆེ། སྐུ་ཉིད་འཛོན་པོ་ཞེ་དགེ་འདུག	[鼓掌]
ང་ལ་གནོད	[开心]
……	……

4.2. The Experimental Procedure

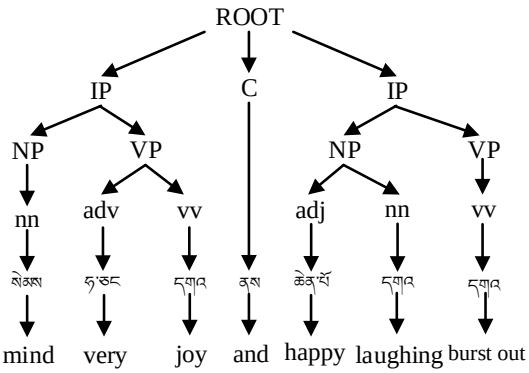
Step1: Parsing the Experimental sentence by the Tibetan parser that developed by Ph.D. Wan FuCheng from China Minorities Information Technology Institute of Northwest University for Nationalities. The Parser takes the parentheses "(", ")" as the unit to construct the syntax tree library. The following example:

Sentence:

ལེམས་ཤ་ཙང་དགའ་ནས་གང་མོ་ཚེན་པོ་དགོད་པ། (My mind very joy and then happy laughing burst out)

Parsing:ROOT(IP(NP(NN(ལེམས mind))VP(Adv(ཤ་ཙང་ very)
VV(དགའ joy))))(C(ནས and))(IP(NP((Adj(ཚེན་པོ happy)NN(གད་མོ laughing))VP(VV(དགོད་པ burst out))))))

Syntax spanning tree:



Step2: Calculate the sentiment value according to the three-layer computing model, and the process follows:

$$V(\text{joy}.v) = 0.5$$
$$V(\text{very}.adv) = 0.25$$
$$V(\text{very joy}.vp) = (|V| + (1 - |V| * |Adv|)) = 0.625$$
$$V(\text{laughing}.n) = 0.125$$
$$V(\text{happy}.adj) = 0.695$$
$$V(\text{happy laughing}.np) = (|N| + (1 - |N|) * |Adj| = 0.733$$
$$V(IP1) = V(\text{very joy}.vp) = 0.625$$
$$V(IP2) = V(\text{happy laughing}.np) = 0.733$$
$$V(sentance) = |IP1| + (1 - |IP1|) * |IP2| = 0.899$$

Calculate the sentiment value of a sentence with these steps and test experimental data and then the experimental results compared with the dictionary method was shown as the Figure4.

Figure 4. The Comparative Data of Experiments

5. Conclusions

Sentiment calculating of the sentence is a challenging and key part of Sentiment analysis. The paper proposed an approach to calculate the sentiment polarity and achieved a good result. Further work is needed to improve the accuracy of the emotional words and to understand the Tibetan grammar rules and try to deal with the Tibetan information from the semantic level.

Acknowledgement

This research has been supported by the Nature Science Foundation of Beijing (No. 4153062), the National Technology Support Program (2014BAK10B03) and the Program for New Century Excellent Talents in University (NCET-12-0579).

References

[1] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Association for Computational Linguistics, vol. 10, **(2002)**, pp. 79-86.
[2] P. D. Turney, "Thumbs Up or Thumbs Down Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of the ACL, **(2002)**, pp. 417-424.
[3] T. Li, T. Zhang and V. Sindhwan, "A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge", Proceedings of the ACL, **(2009)**, pp. 244-252.
[4] A. Abbasi, H. Chen and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums", ACM Transaction on Information Systems, vol. 26, no. 3, pp. 12:1-12:34.
[5] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level", Computer Linguistices, vol. 25, no. 3, pp. 399-433.
[6] P. Melville, W. Gryc and R. D. Larence, "Sentiment Analysis Of Blogs by Combining Lexical Knowledge with Text Classification", The Proceedings of KDD, vol. 9, pp. 1275-1283.
[7] M. Taboada, J. Brooke and M. Stede, "Genre-Based Paragraph Classification for Sentiment Analysis", Processing of SIGDIAL, **(2009)**, pp. 62-70.
[8] Z. Y. Lan and M. J. Zhou, "Semantic Orientation Computing Based on HowNet", Journal of Chinese information processing, vol. 20, no. 1, **(2006)**, pp. 14-20.
[9] L. D. Cheng and Y. T. Fang, "Semantic polarity analysis and opinion mining Chinese review sentences", Journal of computer applications, vol. 20, no. 1, **(2006)**, pp. 2622-2625.
[10] K. Hiroshi, N. Tetsuya and W. Hideo, "Deep Sentiment Analysis Using Machine Translation Technology", Proceedings of COLING, **(2004)**.
[11] T. Zagibalov and J. Carroll, "Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text", Proceedings of COLING, **(2008)**, pp. 1073-1080.
[12] A. Esulia and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining", A Esuli, F Sebastiani - In Proceedings of the 5th Conference on Language Resources and Evaluation, **(2006)**.
[13] S. Baccianella, A. Esuli and F. Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", Proceeding of Lrec, **(2010)**.
[14] P. D. Turney and M. L. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus", National Research Council of Canada, Tech. Rep.: EGB-1094, **(2002)**.

