Micro–blog Emotion Orientation Analysis Algorithm Based on Tibetan and Chinese Mixed Text

Tao Jiang [a], Jing Jiang [b], Yugang Dai [c], Ailing Li [d]

National Languages Information Technology, Northwest University for Nationalities, Lanzhou 730030, China

[a]xinxiyuanjt@126.com, [b]jiangjing_happy@foxmail.com, [c]dyg@xbmu.edu.cn, [d]285962433@qq.com

Keywords: Tibetan Micro-blog; Mixed Text; Sentiment Orientation; POS Sequence; Expectation Cross Entropy.

Abstract. Chinese and English microblog is mostly expressed by single language, nearly 80% Tibetan Microblog presented by Tibetan and Chinese mixed text. If only do emotion orientation analysis of Tibetan or Chinese will lose emotion information, cannot get better effect. By emotion orientation analyzing and studying of Tibetan microblog which is concerned in Sina, making Tibetan Chinese emotion dictionary, Chinese sentences, Tibetan part of speech sequence and emotion symbol as emotion factors and using expected cross entropy combined fuzzy set to do feature selection to realize a kind of microblog emotion orientation analyzing algorithm based on Tibetan and Chinese mixed text. The experimental results showed that the method can obtain better performance in Tibetan and Chinese mixed Microblog orientation analysis.

1. Introduction

As a broadcast social networking platform through attention mechanism share brief information, microblog has attracted a flood of users and has been integrated into people's daily life; publishing and viewing micro-blog has become one of the daily things to do [1]. With the rapid economy development in Tibetan areas and the continuous application of Tibetan information technology achievements, the scale of Tibetan internet users is increasing, the number of micro-blog users are rising rapidly, a growing number of Tibetan users express their feelings and views through the national language on the micro-blog platform. The study of Tibetan micro-blog mining currently is in its infancy, studying Tibetan micro-blog emotion helps analyzing Tibetan short text and micro-blog applied to commercial.

Tibetan micro-blog has large differences compared with Chinese micro-blog, Tibetan micro-blog is more refined in content presentation, a relatively small number of words and relatively concentrated in expressed the theme, and the most significant difference is the diversity of language expression. There are a great number of micro-blog exists Tibetan and Chinese mixing to express semantic, even appearing three languages within a micro-blog, such as Tibetan, Chinese and English. Multilingual mixed express views on the same issue can make a more profound and accurate expression, but also cause ambiguity because of language differences. For example：

“བརྙན་ཐུང་འདི་ལ་གཟིགས་རོགས། ང་ཚོ་ལ་ཕན་ཐོགས་ཡོད་ས་རེད།

Take a look at the video and then think about how to do it.” Its spoken in Tibetan micro-blog is “Please see this video, which is good for us.” Extracted 8000 Tibetan micro-blog in Sina platform randomly and do the statistical analysis, which appeared Chinese and Tibetan mixed expression situation in 6350 micro-blogs, accounting for 79.4% of the total statistical micro-blog. Given the universality of the Tibetan and Chinese micro-blog mixed text in Tibetan micro-blog, micro-blog, taking into account the emotional characteristics of both Tibetan and Chinese languages is necessary while calculation of Tibetan micro-blog emotional orientation analysis.

Micro-blog emotional orientation study is mostly divided into two categories: methods based on emotional dictionaries and methods based on machine learning. Methods based on emotional dictionaries is mainly constructing emotional dictionary or use existing emotional dictionary to calculate the emotional tendencies values by the specific algorithms model, and then classified according to the sentiment value, methods based on machine learning is mainly by building emotional text corpus, changing the texts into a feature vector, and then use the traditional classification algorithms (such as the method of support vector machines, the method of naive Bayes, the method of maximum entropy, etc.) for sentiment classification. The literature [2] built a sentiment analysis system of Twitter, which can do the emotional orientation analysis of the comment information about the presidential election in real time. According to the features of Chinese micro-blog, literature [3] proposed multi-strategies sentiment analysis framework based on the hierarchical structure, and do feature selections to micro-blog properties, such as links, emoticons, emotion dictionary, etc., using a rule based on smileys approach and emotional dictionaries to do emotion classification. Literature[4] proposed a new micro-blog word discovering algorithm to construct the emotional dictionaries based on the emotional dictionary of HowNet, build an automated machine to calculate the short text sentiment orientation after doing the text clause, word, mark, emotional processing. Literature [5] on the basis of the traditional dictionary of Emotional by adding smiley lexicon and new network words to build specialized micro-blog dictionaries, while do rhetorical analysis and sentence analysis micro-blog to effectively improve the effect of tendentious analysis. Literature [6] designed determination algorithm based on the phrase path of micro-blog topic sentiment orientation by building sentiment analysis dictionaries, words online dictionaries and smileys lexicon library for the content features and the forward, reviews relationship features between micro-blog. Literature [7] proposed a theme oriented modelling methods of Chinese micro-blog emotional, the model covers the data pre-processing, parsing, theme extensions, domain knowledge, polarity adjustment in the context of emotional word, smileys, and more.

The research on micro-blog sentiment analysis mostly focused on Chinese and English micro-blog currently, and Chinese micro-blog is published in one language, but for Tibetan micro-blog sentiment analysis research especially for Tibetan and Chinese mixed micro-blog sentiment orientation study has not been reported. From Tibetan micro-blog particularity of Tibetan and Chinese mixed, combining emotional feature with emotional expression of Tibetan and Chinese micro-blog text put forward emotions tendentious analysis algorithms with the Tibetan and Chinese micro-blog mixed text.

2. The micro-blog sentiment features

Sentiment lexicon. Sentiment words are the words contained the emotional information, which express the emotion of the inner, emotional words can express the sentiment orientation, so using the polar of the emotional words in text can promote the effects of evaluating. The objects processing in this paper is Tibetan and Chinese mixed texts, so both of the Chinese sentiment lexicon and the Tibetan sentiment lexicon are been set. The Chinese sentiment lexicon is based on HowNet and NTUSD, HowNet has published “sentiment analysis word set”, including “Chinese sentiment analysis lexicon” and “English sentiment analysis lexicon”, there are approximately 8942 Chinese words；NTUSD is a Chinese sentiment lexicon sorted out by Taiwan University, there are two versions of Simplified Chinese and Traditional Chinese, 2812 positive sentiment words included in each version, and 8276 negative sentiment words. There are no basic sentiment lexicon in Tibetan language, automatically extracting adjectives, verbs and nouns as the candidate items of emotional words from a large number of Tibetan micro-blog in this paper, selecting the words with strong emotion as reference words artificially, calculating the similarity of candidate words and reference words by the method of so-PMI to judge the sentiment orientation of candidate words, and then embody the words with more sentiment orientation to the lexicon[8].

Chinese Sentence features. Micro-blog texts almost are various sentence forms, strong arbitrary, the major sentence are exclamatory sentence, rhetorical sentence, interrogative sentence, declarative sentences, etc., the degree of different sentences, the sentiment reflected are different. Exclamatory sentence is to express strong feelings and express sigh tone sentences. Exclamatory sentence interjects often with "many", "how", "too", "rather", "good", "very" and other adverbs. Exclamatory sentence is the strongest sentence to express sentiment in Chines. Rhetorical sentence actually are emphasizing on certain positive or negative answers of commentators, which is the emphatic performance of declarative. Interrogative sentence express a state of mind questions on objects of commentators, commentators do not understand but concern about that. In addition to declarative sentences, sentiment factor of interrogative sentence is the lowest. According to each sentence expressing sentiment features, Chinese sentence will be the feature and combine different punctuation given different sentence to the appropriate sentiment factor.

Tibetan POS sequence features. As the grammar attributes of words, POS reflect the role words play in the sentence, POS can solve the ambiguity between polysemy; Therefore, POS can be used as sentiment orientation features analysis. To negative words, when negative words appeared in the sentence will change sentiment elements existing and directly affect sentiment intensity of the sentence. POS tagging of Tibetan micro-blog by "Tibetan parts of speech classification and labelling set" proposed by Professor Qi Kun Yu in Northwest University for Nationalities in this paper, the tagging set contains 21 POS categories and 61 symbols, which have done a detailed for each part of POS classes, To adverbs is the degree adverbs (ཤིན་ཏུ་), negative adverbs (▯ ▯), certain adverbs(ཁོ་ཐག)etc. [9].

Micro-blog Symbols features. Smileys in microblog text can express emotion more simple and intuitive which have a very important role on reflecting the author's sentiment, the user can express the attitude through smileys that are difficult to express in words. As the positive expression of 🤗is the meaning of “[applause]”, Which is Said the publisher is strongly agree of the things; 👍is the meaning of“[great]”, which is Said the publisher intended to praise the subject of the comment; the negative expressions as 🤬 is the meaning of "[Curse]", the author used to express hate something; 😫 is the meaning of “[crazy]”, expressing angry but no place to vent, reflecting the negative emotions of publisher. Tibetan text microblog texts mainly from Sina microblog and Tencent microblog, building smileys library based on the microblog expressions, classifying the smileys into positive strong sentiment, positive ordinary sentiment, negative strong sentiment and negative ordinary sentiment in accordance with the direction and extent of sentiment, and giving the corresponding sentiment values to the two levels of sentiment smileys of "strong" and "Ordinary ".

3. Representation model of sentiment

To represent the sentiment features for the Tibetan and Chinese mixed text the Vector Space Model is adopted in this paper, combining expectation cross entropy with fuzzy set to do sentiment selection to select the feature which can express sentiment orientation more clearly and solve the problem of sentiment ambiguity. Expectation cross entropy always consider the relationship between the data features and categories, which is not only can analyse the appearance of sentiment words impact on classification better, but also can eliminate ambiguity by training data, and fuzzy set for correcting sentiment words, which can evaluate the accuracy of words and expressions uniformly, and enhance the accuracy of sentiment classification.

Expectation Cross Entropy. Expectation cross entropy is also called relative entropy, which is a parameter estimation method based on information theory. Its original meaning is: when the true distribution h(x) of x is not known, supposing x obey the distribution of g(x), and then calculating the distance between g(x) and h(x) which is the cross entropy[10]. The distance between g(x) and h(x) is also called distance KL, which is the abbreviation of Kullback-Leibler Divergence, the formula is as follows:

$$\mathrm{Dist}(g,h) = E_g \log\frac{g(x)}{h(x)} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)\log\frac{g(x_i)}{h(x_i)} \tag{1}$$

The principle of expectation cross entropy is the same with information gain, the only difference is that expectation cross entropy is not considering the case of features do not appear. It reflects the distance that the probability distribution of text categories and the probability distribution of text category under some feature item, the larger expectation cross entropy of feature item, the larger impact on the distribution of text category.

Sentiment features correction of fuzzy set. The core of fuzzy set is subordinate function. The relative of sentiment feature and sentiment categories in sentiment classification can be expressed as a fuzzy map: R:T×C→[0,1], $\mu_R(t_i, C_j)$expressed in the map R, the membership degree sentiment feature t_i to the sentiment categories C_j, μ_R is the subordinate function[11]. The subordinate function is described the extent of the feature belonging to fuzzy set. The value of membership degree is determined by calculating the training text, w_t is expressed the microblog which has appeared sentiment words t_i , $c(w_t) \in C_j$ is expressed that microblog w_t belongs to sentiment categoriesC_j.The membership degree formula of sentiment words to sentiment categories is:

$$\mu_R(t_i, C_j) = \frac{\sum_{t_i \in w_t,\ c(w_t) \in C_j} w_t}{\sum_{t_i \in w_t} w_t} \tag{2}$$

If a sentiment word appears only once in the text of a category, then the membership degree value is 1,which indicating that the sentiment word has absolute classification effect on the sentiment categories, if an sentiment word appears in more than one categories, then the sentiment category membership values of the word is determined by the number of micro-blog belonging to the sentiment category divided by the number of text containing the sentiment word of micro-blogs, the direction and extent of the sentiment words is more clear after the introduction of this method.

4. Experimental Results and Analysis

Experimental corpus. Using the information of Sina micro-blog as experimental corpus, using the open platform interface of Sina micro-blog to crawl 2, 5000 Tibetan micro-blog, remove duplication and pure links micro-blog, 15,000 micro-blog corpus were selected as the experimental corpus. Experimental corpus is divided into two parts, the first part of the corpus is mainly Tibetan content, containing a small amount of Chinese or do not contain Chinese at all, the second part of the corpus is Tibetan and Chinese mixed text, the amount of bilingual content are more balanced. Detailed experimental data are shown in Table 1:

Table 1 Corpus type and proportion of Tibetan micro-blog

Corpus	Number	Sample
Tibetan	2000	གནས་ལུགས་མང་པོ་ཞིག་ཤེས་ནའང་། མི་ཚེ་ནི་སྔར་བཞིན་འཁྱོངས་བར་དཀའ་ཡང་།
Mixed Text	12000	ཁ་སང་མགོ་ལོག་ཁུལ་མི་རིགས་འཛིང་གི་སློབ་མ་བདུན་བརྒྱ ་ལྷག་གིས་དུ་ཆང་བཅད་པས་ཁོ་ཚོའི་བསམ་དོན་འགྲུབ་པར་སྨོན། Today, more than seven hundred students of Golog national high school quit drinking and smoking really touched people excited.

Corpus pre-processing. Before experiment, the corpus must be preprocessing as the following steps:

(1)Filtration. Remove the URL link, the user name (such as @ YYY, @ user ID), topic (The topic in Sina micro-blog is formatted as # XXX #).

(2)Language identification. Experiments corpus involving Tibetan and Chinese two mixed texts, so there need to distinguish for different languages texts, combining text encoding with high-frequency words to releasing language identification.

(3)Segmentation and POS tagging. Chinese text using ICTCLAS to do segmentation from Institute of computing technology, Chinese academy of sciences. Tibetan word segmentation and

POS tagging using HMM-based Tibetan POS automatic annotation software developed by Professor Qi Kun Yu in Northwest University for Nationalities.

Experimental Analysis. (1) Experiment between Tibetan-Chinese mixed feature and monolingual features. This experiment was validated that do orientation in the situation of Tibetan - Chinese mixed features is superior to monolingual orientation features by using the same corpus. Exacting randomly 5000 Tibetan and Chinese mixed micro-blogs as the experimental corpus, text representation models are all using the vector space model, using the expectation cross entropy as the method of feature selection, using SVM algorithm library Libsvm developed by Professor Chih-Jen Lin from Taiwan University to training and classification of sentiment analysis model. Extracted respectively Tibetan features, Chinese features, Tibetan and Chinese features to do three sets of experiment, results are shown in Table 2.

Table 2: Sentiment classification results

Feature type	P	R	F-value
Tibetan feature	0.590	0.600	0.595
Chinese feature	0.691	0.713	0.702
Tibetan and Chinese mixed feature	0.819	0.845	0.832

Tibetan most expressed Buddhism and religion information included in Tibetan and Chinese mixed micro-blogs, extracted Tibetan features only to distinguish sentiment categories is difficult, so the sentiment classification results are worse, the part of Chinese is doing the statement of facts in Tibetan and Chinese mixed micro-blogs, the sentiment orientation which reflected is not obvious, although the experiments effect of extracting Chinese text feature to do sentiment classification are better than the Tibetan features, but the accuracy rate does not meet the 70 percent, the effect of extracted Tibetan-Chinese mixed texts sentiment features to do sentiment classification is much better than monolingual features, compared with extracting Chinese features only, accuracy of classification is increasing by 13 percent.

(2) Experiment of different feature selection method. To validate the effectiveness of expectation cross entropy combined with fuzzy set correction method proposed in this paper, selected 2000 randomly Tibetan and Chinese micro-blog mixed corpus to do comparative experiments of text representation, the text representation method are all using vector space model, feature using expectation cross entropy, information gain and mutual information as the selection methods, experimental results are shown in Table 4.As the experiment results shown that the accuracy rate of method proposed in this paper is higher 4% and 5% than respectively information gain and mutual information, for the method of expectation cross entropy combined with fuzzy set is more concerned about the relationship between sentiment words and sentiment categories, and using the degree of membership values to quantify, the information gain method is very sensitive to sentiment words appears very few, when dealing with short text the efficiency will be reduced.

Table 3: The results of different feature selection

Feature selection	Accuracy rate	Recall rate	F value
Expectation Cross Entropy	0.828	0.841	0.834
Information Gain	0.794	0.812	0.803
Mutual information	0.774	0.791	0.782

5. Conclusions

As a mainstream communications tool, micro-blog has a strong social influence, research on Tibetan micro-blog emotional sentiment orientation analysis is in favor of the development of Tibetan public opinion techniques. In this paper proposed a Tibetan micro-blog sentiment orientation analysis algorithm based on Tibetan and Chinese mixed texts aiming at the prevailing situation of Tibetan and Chinese mixed text in Tibetan micro-blog, extracting the sentiment features from the semantic point of view, combining the fuzzy sets with expectation cross entropy to do the feature selection. The orientation analysis experimental based on bilingual feature was superior to the monolingual features. The next step will do orientation analysis experimental more effectively by integrating of semantic features, expanding and optimizing Tibetan sentiment knowledge base.

Acknowledgment

This research has been supported by the program of central specialized funds financial aid for postgraduate of Northwest University for Nationalities (Yxm2014001), the project of National Natural Science Foundation “Study of Semantic Relationship Extraction of Tibetan Entities (61262054)”, and key project of science and technology of Gansu Grant No. 1203FKDA033.

References

[1] Wen kunmei, Xu shuai, Li Ruixuan, et al. Survey of micro-blog and Chinese micro-blog information processing [J].Journal of Chinese Information Processing, 2012, 26(6):27-37.

[2] Hao Wang,Dogan Can, Abe Kazemzadeh. A system for real-time Twitter sentiment analysis of 2012 U. S.presidential election cycle[C]. Proceedings of the ACL 2012 System Demonstrations, 2012: 115-120.

[3] Xie Lixing, Zhou Ming, Sun Maosong. Hierarchical structure based hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction [J].Journal of Chinese Information Processing, 2012, 26(1):73-83.

[4] Han zhongming, Zhang Yusha, Zhang Hui, et al. On effective short text tendency classification algorithm for Chinese micro-blogging [J]. Computer Application and Software, 2012, 29(10): 89-93.

[5] Liu Peiyu, Zhang Yanhui, Zhu Zhenfang, et al. Micro-blog orientation analysis based on emotion symbol [J]. Journal of Shandong University (Natural Science), 2014, 49(11):8-13.

[6] Liu Quanchao, Huang Heyang, Feng Chong. Multi-feature sentiment orientation identification algorithm for micro-blog Topics [J].Journal of Chinese Information Processing, 2014, 28(4):123-131.

[7] Lu Hao, Niu Zhendong,et al.A Model for Sentiment Classification of Chinese Micro-blog Based on Parsing and Theme Extension[J],Tansactions of Beijing Institute of Techonology, 2014, 34(8): 824-830.

[8] Alina Andreevskaia, Sabine Bergler. Mining WordNet for a fuzzy sentiment: sentiment tag extraction from WordNet Glosses[C].Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento: Association for Computational Linguistics, 2006: 209-216.

[9] Lu Weisheng, Guo Gongde, Chen Lifei. Emotion classification with feature extraction based on part of speech tagging sequences in micro blog [J]. Journal of Computer Applications, 2014, 34(10):2869-2873

[10] Pu Qiang , Yang Guo Wei .Short-Text Classification Based onICA and LSA[J] // Proceedings of International Symposium on Neural Networks, 2006(ISNN 2):265-270.

[11] Fu Xuefeng, Wang Wenming. Text Categorization Method Based on the Theory of Fuzzy-rough Sets [J].Journal of South China University of Technology (Natural Science Edition), 2004, 34(32): 73-76