Research on Tibetan Text Orientation Identification

Xiaodong Yan
Minzu University of China, Beijing Haidian district zhongguancun street 27#, 100081, China
National language resource monitoring &Research Center Minority Languages Branch
E-mail: yanxd3244@sohu.com

Xiaobing Zhao
Minzu University of China, Beijing Haidian district zhongguancun street 27#, 100081, China
National language resource monitoring &Research Center Minority Languages Branch
Email:nmzxb@163.com

***Abstract*— In recent years, Minority languages in China are widely used on the computer and network. But now there is no effective public opinion analysis system of the minorities overall attitude of the masses of the hot events or topics. In this study, we research on Tibetan topic orientation recognition. First, according to the Tibetan context and life characteristics, combined with a set of emotional words in Hownet, the Tibetan emotional word dictionary is built, and then by the Tibetan word semantic similarity calculation method we extend this dictionary to get rich emotional word set. We also propose a method that the sentence orientation is determined by the orientation of words in this sentence and the orientation of text is determined by the orientation of sentences in this text. By our research the Tibetan hotspot information can be rapidly detected and found and then the public opinion tend can be track quickly. It is benefit for positive guidance of public opinion.**

***Index Terms*—orientation recognition, semantic ontology, emotional dictionary**

I. INTRODUCTION

With Web2.0 era coming, network has gradually become an important carrier of public opinion. Discovery of public opinion from network and excavation of Internet users' view and tendencies also become a new hotspot. In recent year Minority languages in China are widely used on the computer and network, but there is no effective public opinion analysis system to express people's attitude on the hot events and topic. In this study, we research on Tibetan topic orientation recognition.

Analysis of emotional tendency can be roughly divided into word tendency analysis, sentence sentiment orientation analysis, text emotion tendentious analysis, the overall bias prediction of mass information four level studies [1].

Word tendency analysis includes words polarity, intensity and context mode analysis. The processing target of word tendency analysis is a single word or entity and the processing target of sentence sentiment orientation analysis is the sentence in the specific context of the statement. Text emotion tendentious analysis is the overall judgment on the emotional tendencies of a text. The overall bias prediction of mass information is for huge amounts of data. Its main task is to extract information which is all about one certain topic from different sources. Then all the information will be integrated and analyzed in order to dig out the characteristics and trends attitude.

Text orientation analysis techniques are applied in network Business Review, online public opinion analysis, network filtering and other fields. People can be guided by the evaluation online when they buy a product. The technology can also help us monitor network public opinion and so on. Common text orientation analysis methods mainly include statistical machine learning method and semantic-based approach.

Machine learning is a method of Learning from Data.

In it machine learning algorithms are used for statistical language model training, and then the new text are recognized by use of the trained classifier. In 2002, Bo Pang et al. first introduced machine learning methods in the field of emotion analysis [2]. They use Native Bayes, Maximum Entropy and Support Vector Machines classification methods to classify the documents on the document level. In 2004, Bo Pang etc. also proposed through machine learning and graph min-cut method to judge for the sentences of documents [3]. In 2007, Ni etc. used the CHI and information gain for feature selection and use NB, SVM and Rocchio's algorithm for emotion classification [4] . In 2006, Cui etc. used PA (Passive Aggressive), LM (Language Modeling) and Winnow classifier for emotion classification and compared their performance [5]. On using machine learning methods for Chinese emotional analysis, in 2005, Ye Qiang etc. [6-7] extracted from the text subjective information and gave appropriate weight, and then built tendencies classifier based on the weight. Cai Jianping etc.[8]proposed the

The work in this paper is supported by the National Natural Science Foundation of China project (61331013).
Author: Xiaodong Yan,email:yanxd3244@sina.com

doi:10.4304/jcp.9.7.1598-1605

methods for judging the orientation of words and sentence based on machine learning.

The Semantic-based text orientation research method is mainly based on emotional tendencies dictionary which is generated by expanding electronic dictionaries or word knowledge base. As Zhu Yan et al [9] use semantic similarity and semantic relevance computing to provide the relevance of pre-selected words and base words. Turney in his paper [10] also introduces the unsupervised text classification method based on semantic orientation. In 2004, Hu[11] first proposed the application of association rules for extract the product characteristics of English reviews. By use of this unsupervised method for mining the mobile phones, digital cameras and other product reviews. The average recall rate is about 80% and the average precision ate is of 72%. And the follow-up study [12] can help people determine the user's emotional guidance on these characteristics. Kobayashi etc. [13] used a semi-automated circulation method to extract product features and user opinion.

For text orientation research, Ye [14] explored the Chinese document sentiment analysis theory and proposed a Chinese emotional semantic analysis method based on PMI-IR method. This method can obtain analytical results of similar studies in English, showing that the method in Chinese sentiment analysis on application prospects. He tingting [15] improved the text classification method based on HowNet semantic similarity calculation to determine the tendency of text emotion.

Currently text orientation analysis and mining research on minority language is little. The investigation has not found any research achievements on Tibetan text orientation mining.

II. The Structure of Tibetan Text Orientation Analysis System

Forum Web site is a specific form of the general Speaking forum page in which there contains three main parts:

(1) web format information, website information , managers and User links , advertising links, etc. ;
(2) Poster, post respondents and their names, time and location of landing or publish and other related information ;
(3) The context of the post and the context of the post respondents.

In our research, we only concerned section (3). So we will filter unwanted marking and information to extract the context of post and the each respondent's corresponding content of reply the post.

From the grammatical forms, we find that there are multiple information units in the web page. The arrange of them is compact and the style of them are similar. You can regard every reply as an information block. So that tags can be based on the HTML file to create analysis tree to determine the information block.

We have analyzed the web tree and find that the forum page file has following characteristics [16]:

1) There are multiple information blocks in the Web file and these information blocks unified a whole unit. The whole unit is in a discrete area of the Web page.
2) <TITLE> and </TITLE> are used to tag every theme. Each information block is then used to describe similar tags, and a plurality of information blocks is at the same level of the tree.

Based on the above analysis of the pages we can extract specific text as HTML markup.

Judgments of the web is judged based on the tendency of a single document which on the chapter level and in natural language processing area, understanding of chapter is based on the understanding of the paragraph, comprehension of the paragraph is based on the sentence or sentence group. Then sentence comprehension due to the understanding of words. So orientation research on chapter is also based on the tendency of word tendency research and at the same time the internal grammatical structure of sentence and the relationship of sentences and paragraphs are should be considered.

According to the particularity of the Forum web page, the web text is divided into a number of pieces of information after the pretreatment. The forum theme is the base and one information block can be decomposed into a number of sentences. One sentence is expressed by a number of characteristic words. So the judgment of text can be shown as figure 1.

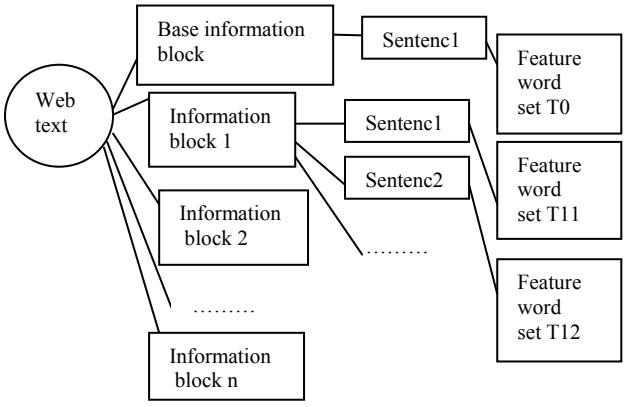

Figure 1. Web text orientation judge method

We proposed a Tibetan text orientation recognition method just based on above analysis, as shown in Figure 2. And it will work as three aspects.

A. Constructing the Tibetan Emotional Dictionary

Currently, in China the most authoritative emotional word thesaurus is the "words set for sentiment analysis" provided by HowNet, but there is no emotional intensity in it and still needs to be improved. Moreover constructing the emotional Tibetan dictionary is not just a simple translation work from Chinese to Tibetan, but according to the Tibetan grammar and linguistic characteristics, then we can build the emotion dictionary for Tibetan orientation analysis work.

In our study, we will refer to the emotional words set published by HowNet.

B.Computing the Semantic Orientation of the Tibetan Words.

Computing the Semantic Orientation of the Tibetan words is fundamental to building emotional dictionary, but also it is the basis of lexical semantic polarity judgment. Research on the orientation of words is the premise of the research on orientation of text. words with emotional tendencies are mainly nouns, verbs, adjectives and adverbs but also include the name, organization name, product name, event name and other named entities. Among them, some of the judgments (otherwise known as polarity, usually divided into compliment, derogatory and neutral three kinds) of words can be obtained through the dictionary; polarity of the rest of the words cannot be obtained directly and must be obtained by the semantic computing. The polarity of emotional tendencies of words is not including polarity but also the tendency extent. This kind of judgment requires a combination of adverbs solutions.

Referring to the semantic similarity algorithm on Chinese vocabulary [9], we proposed a Tibetan word similarity algorithm based on semantic. And we also designed to build a Tibetan modifier dictionary.

C. Identifying the Tibetan Text Orientation.

After computing the Semantic Orientation of the Tibetan words, we can do the sentence level emotion tendencies recognition work. Judgments for words are to process single word or entity of the sentence, but the subject of text orientation analysis is the specific context of the sentence. Its mission is to analyze and extract the various subjective information in the sentence, including the judgment of sentence emotional tendencies, as well as extract the discourse and various elements associated with emotional tendencies, including the holder of emotional tendencies and the evaluation object, tend polarity, strength, and even the importance of discourse itself.

Chapter -level emotional orientation analysis is a overall judgment of a text.

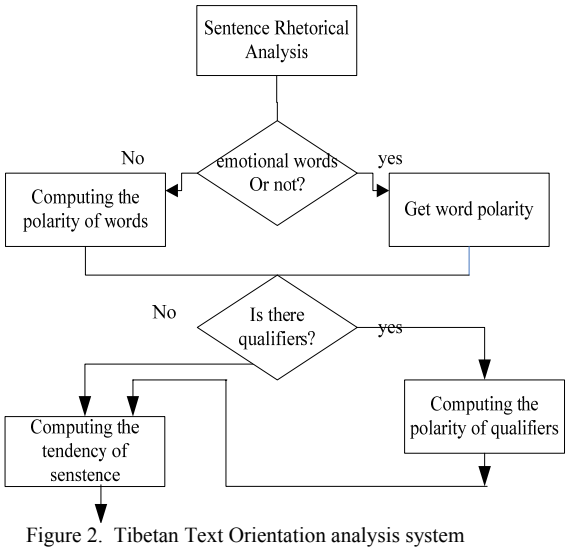

Figure 2. Tibetan Text Orientation analysis system

For Tibetan text recognition work, we first have the word emotional tendencies in a sentence and by use of them to judge the polarity of this sentence. At last we can get the bias of the whole text through the value of the cumulative sentence emotional tendencies.

III. TIBETAN EMOTIONAL DICTIONARY

Currently, the most authoritative emotional word thesaurus is the "words set used for sentiment analysis" in HowNet, but there is no emotional intensity in it and it still need to be improved. For constructing the Tibetan emotional dictionary is not simply translating Chinese to Tibetan but according to Tibetan grammar and linguistic characteristics, building a Tibetan emotional dictionary for Tibetan text orientation analysis work.

In our study, we intend to build a Tibetan emotional dictionary based on HowNet emotional words set, according to Tibetan grammatical features and Tibetan features of life.

A. Selecting Basic Emotion Words

On the basis of emotional words in HowNet, and according to the Tibetan grammatical features, we build the basic Tibetan emotional word Thesaurus. There are about 6000 Tibetan emotional words in it. Commendatory terms and derogatory terms are each about 3000.

B. Selecting Seed Words

The basic emotion words are sorted by the number of hits from Google search and the highest number Hits words are selected. According to the reference of [17], the number of seed words is about 15% of the total emotional words and in this case, the accuracy of emotional orientation judging is about 90%.

The accuracy is also stable. Therefore, we will determine the seeds of about 900 words, including positive comments seed words 490, and negative comments seed words 410. For example, we select some seeds of positive and words as following:

ཨཁཁས་ (beautiful), དངར་བ་ (popular), དགཔོཙ་(positive) , ཆ་ཤ་གསལ་བ་ (perfect) , དཔཁན་ཙོཁ་ (good), excellent (ཡང་), ཁང་བ་ (boring), ལཁསལ་ངས་ (health) , ཁཏ་འབད་ཙཁ་ (wonderful) , དཀ་ (poor), གས་ད་གག་ (lonely), དབག་པ་ (real), འཚ་པ་ (convenient), དར་ལ་ (fashion), གཁཏ་འ་རང་དགའ་ (happy), བཀའ་ཆ་ (thank), དཀའ་གས་བ་ (hard).

The negative seed words as following:
་བ་ (desolate), འལ་བ་ (error), དངངས་ག་ (terror), ངན་ཚ་ཤ་ཁ་ (bad), དགའ་བ་ (like), གཁྱན་འཇགས་ང་བ་ད་པ་ (stable), ང་ (old), དཀ་པ་ (depressed), ཆག་འཁས་ (dark), ཨས་བངང་ཨཆན་པ་ (illegal), ལས་ན་པ་(negative), ཇངང་(confusion), སྤྱོད་ཚོར་ (waste), ཁཆ་ཆ་ (crazy) , དུག་ (poison) , དཀའ་ལས་ག་(difficult) , ཁཆ་པ་ (weak), ཤབས་ད་(helpless),
ལ་ (smelly).

C. Calculating the Emotional Tendencies Weights of Base Emotion Words to Achieve Polarity Judgment

For two words w1, w2, assuming that there is multiple meaning of the original for each word, w1 is p11... p1n, w2 is p21... p2m. The word similarity calculation formula of w1 and w2 is as following:

$$sim(w_1, w_2) = \max\left(sim(p_{1i}, p_{2j})\right); \quad 1 \le i \le n, 1 \le j \le m \tag{1}$$

In it: $sim(p_{1i}, p_{2j}) = a/(d + a)$ is the semantic distance between the two words. d is the route distance between the two word. a is an adjustable parameter.

The emotional weight of the word w is determined by the similarity degree of w and each word of seed set.

Assuming seed set = {PP, PN}, PP refers to compliment seed word set, PN is derogatory seed word set, and then the emotional weight of word w is defined as following:

$$o(w) = \frac{1}{M}\sum_{i=1}^{M} \mathrm{sim}(w, pp_i) - \frac{1}{N}\sum_{j=1}^{N} \mathrm{sim}(w, pn_j) \tag{2}$$

Where: ppi ∈ PP, pnj ∈ PN, M and N is respectively the number of compliment seed set and derogatory seed set. 0 is set as the threshold.

Here o(w)>0 indicates that the term is commendatory. o(w)<0 indicates that the term is derogatory. The value of o(w) represents the emotional degree of word w.

D. Constructing Base Emotional Word Dictionary

According to the constraint domain principle, the polarity of emotional word is confined in a closed interval, to facilitate quantitative analysis.

We use [-1, +1], a symmetric interval to identify the emotional tendencies and polarity degree of words. “0”represents neutral emotion. Positive territory represents compliment tendency and negative territory represents derogatory tendency. Larger the absolute value is, stronger the emotion is. We use the linear method re-planning the emotional weight. Specific formula as formula (3):

$$d' = \frac{d - d_{min}}{d_{max} - d_{min}} \tag{3}$$

In it d' is the re-planning emotional weight. d is based on the formula (2). d_{min} is the minimum weight of the weights calculated from formula (2)and d_{max} is the maximum value. We need calculate the emotional weights of all the words according to equation (3) and remove the incorrect classification words and not helpful words. At last we can obtain the emotional dictionary in which there are positive emotion words, negative emotion words and neutral words. Table 1 is the weights of part emotional words.

TABLE I.
WEIGHTS OF PART EMOTIONAL WORDS

positive words	weights of Positive words	Negative words	weights of Negative words
ཨཁཁས་ (pretty)	0.988	འཁག་ (disorder)	-0.981
ད་འཁཁག་ (better)	0.918	ག་ལ་ད་པ་ (stolid)	-0.927
ཆ་ཤ་གས་པ་ (excellent)	0.913	ཁང་ག་ཚ་ (crude)	-0.864
འཁངང་ས་ (first level)	0.830	ཆག་ཡོ་ (dark)	-0.846
ཁ་དག་ (refined)	0.824	འག་ང་ (dirty)	-0.826
དགའ་ཤ་ (happy)	0.821	ངས་གས་ (arrogant)	-0.796
ང་ཤ་པ་ (comfortable)	0.819	ངན་ཚ་ཤ་ (bad)	-0.700
ཆ་ང་འཁས་པ་ (elegant)	0.813	ཨཁས་གག་པ་ (terrible)	-0.693
ཐྲིད་ཚོ་ (warm)	0.795	འཆ་ཨངང་ད་པ་ (not good)	-0.677
འཁ་ཚཁག་པ་ (great)	0.881	ཆ་ག་པ་ (disgusting)	-0.858

E. Constructing Tibetan Adverb Dictionary

There are usually a lot of adverb words in text, such as “more important”, “very unfair”. “More” and “very” are strong in emotional express; “no” is also used in sentence and these negative words and adverbs usually change the polarity of the original word. In order to better analyze the semantics of the emotional word, we must first analyze the adverb and negation word. So we will construct adverb dictionary. For example:

very (གང་འཚཚ་གཁ་), more (ཇང་ང་) and so on.

According to the reference [18], we selected the commonly used adverbs vocabulary, as shown in table 2.

TABLE II.
THE STRENGTH DEGREE OF COMMONLY USED ADVERBS

Adverb	strength
དག་གནན་(the most)	1.0
དག་གནན་(very)	0.8
ལག་(better)	0.2
ང་ཙམ་(a little)	-0.7
ང་ཙང་(little)	-0.5
ཚད་བར་(seldom)	-0.3

IV. TIBETAN VOCABULARY POLARITY JUDGMENT

After the base emotional words are gotten by semantic similarity calculation, we will expand the base emotional dictionary by automatic emotional words polarity judgment method. We refer to [19][20][21][22][23][24]The method as following:

1) First the emotional word dictionary is searched for every candidate emotional word. If there is the word we searched in the dictionary, we could get its polarity and strength;

2) If this word is not searched, we should search the emotional words forward and backward and find the related words with these words;

3) If there is no related word, the polarity of this word is calculated by the formula (2).

4) If there are related words between the candidate emotional words and the front or back of this emotional word. We will first determine the type of the related words, and then calculate the polarity and strength according to the related word type.

V. TIBETAN TEXT ORIENTATION RECOGNITION

In order to calculate text tendentious, we will search emotional word w in the sentence according to the emotion dictionary. We record the tendency of word w, o (w). Then the emotional tendency values of all the emotional words are accumulated and the emotional tendencies of the text are given. Suppose that there are n sentences in a text, sen_1, sen_2... sen_n, and there are k emotional words in sentence sen_m,w_{m1}, w_{m2}, ..., w_{mk}. The tendency of the entire text as formula (4):

$$D = \sum_{i=1}^{n} \sum_{j=1}^{k} o(w_{ij}) \quad (4)$$

Through the above work, the Tibetan text orientation can be achieved. But in text orientation there are also some problems to solve. They are negation sentence and sentence including adverbs.

A.Negation Sentence

In logical semantics, Negative word is the body of judgment and it does not have certain characteristics or behavior. For example:

His acting is not progressed.

He is not confident on his performing .

In the two sentences, progress and confident are both negative words. But when “not” or “no” appeared in the sentence, the semantics of the whole sentence will change to a pejorative one. So we need solve these problems. We use negative rule match method. The orientation of the word which is matched will be reversed in order to correctly reflect the view of the entire corpus. First, we will extract negation from our corpus, and then in a large number of negative sentences we should extract high-frequency negative rule set. At last we can compare the negation rules with the negation sentence. If the negation center is the words with emotion tendency, it will be replaced with the opposite meaning of the word in order to eliminate the effect of negative sentence for the text view. In this article we achieve the negative word referring to HowNet. We will select the sememe with negative meanings, such as: {negative}, {Be unable}, {impossible}, {unable}, etc., and extract the original concept containing negation sememe , then determine the negative words by artificial filtration.

B. Degree Adverbs Semantic Strength

As we described in section II there are usually a lot of adverb words in text, such as “more important”, “very unfair”. “More” and “very” are strong in emotional express in sentence and some adverbs usually change the polarity of the original word. Adverbs are always having different strength on emotion express. Either absolute or relative degree adverbs of the sentence will have huge impact strength on the text semantic. For example:

His Chinese is good.

His Chinese is very good.

His Chinese is absolutely good.

Semantic strength of the three sentences is in ascending order. In order to better distinguish the emotion strength of the judgment view, in this paper we set up a viewing window for adverbs. The size of the window is a parameter derived from the training set which is the best option. Here the window size is calculated based on the extent of the word with the degree adverbs, not the number of words they are apart. If the orientation words appear in the observation window, the frequency of words is increased by the level of the degree adverbs.

VI. EXPERIMENTS RESULTS AND ANALYSIS

In our experiments blog collection is downloaded from the website http://blog.amdotibet.cn/(it is the Qinghai lake blog website which is built in Qinghai province and many Tibetan people use it in them language). We select about 100 blogs from the same subject to judge the polarity of the blog by our method.

In our experiments, we will respectively do several jobs including feature words extracted, web information block text orientation judgment and judgments on web text orientation. We will test the results by manual inspection to verify the validity of our method.

The experimental procedure is as following:

Step1: Collecting a certain amount of web forums and preprocessing them to get Topics and reply part text, then the initial corpus is formed;

Step2: For each text, do the following treatments:

Step2.1: Manually extracting a small amount of judgment benchmark words, and were saved as collection WP and WN;

Step2.2: For each sentence, by use of the Tibetan segmentation and labeling software developed by our research center (National language Resource Monitoring & Research Center Minority Languages Branch), we achieve sub-word annotation and syntactic analysis, extraction labeled as nouns, adjectives, verbs. According to the formula (3), the judgments of the word are achieved, and they are added to WP or WN collection.

Step2.3: According to the formula (4), the polarity of

text is calculated.

Step3: Outputting the judgments feature words extracted from web and giving the tendency of the current web page.

The judgments feature words extracted from web can be test manually. In the experiments, there are three methods to manually select judgments words.

Five derogatory terms are selected in the first group; five commendatory terms are selected in the second group; Five derogatory and five commendatory terms are selected in the third group;

The terms of three groups are as input experiments data to verify our research.

Table 3 is the results of the three experiments. It is the artificial proofreading results. Figure 3 shows the data of table 3 in graph.

TABLE III.
THE JUDGMENTS CHARACTER WORDS OF THREE TEST TIMES

Test method	Word number correctly extracted	Word number wrong extracted	Word number no extracted
The first group	132	35	34
The second group	123	31	45
The third group	110	67	23

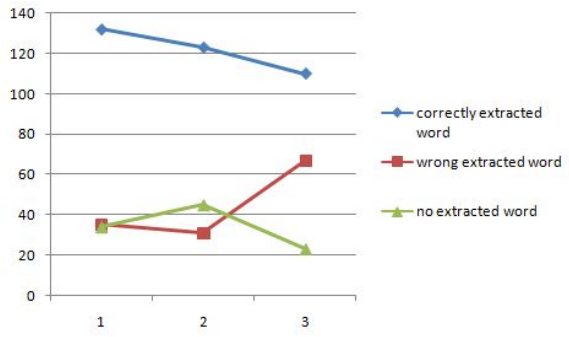

Figure 3. Comparison chart of three tests

We define the word correctly extracted rate is that the number of correctly extracted words is divided by the total number of words and the omission rate is that no extracted word number is divided by the number of whole emotional words. Then the word correctly extracted rate of the above three methods are 0.79, 0.8, 0.62 respectively. The omission rate of the above three methods are 0.17, 0.23, 0.12 respectively.

From the above test results we can see that when we use only one tendency words the correct rate is higher than we use both derogatory and commendatory words. But when we use both derogatory and commendatory words in the test we can obtain higher missing rate than we use only use one tendency words.

We use our method to calculate the orientation of text, and compare the results of our methods with artificial evaluation value. We define that the difference between the two values is the evaluation deviation.

We evaluate our research results on single sentence, information blocks, and web page text respectively and calculate the evaluation deviation respectively. In our experiments we select 100 sentences, 100 information blocks and 250 webpage text and all the data are examined. We obtain the arithmetic mean of evaluation deviation values and they are shown in Table 4. Figure 4 shows the data of table 3 in graph.

TABLE IV.
THE ARITHMETIC MEAN OF EVALUATION DEVIATION VALUES

Evaluation object	sentences	Information blocks	Web pages
Evaluation deviation	0.27	0.45	0.30

Because when we extract the derogatory and commendatory words and do similarity calculation, we always use syntactic analysis. We judge the tendency of sentences by use of the extracted derogatory and commendatory words can obtain high correct rate and small evaluation deviation value.

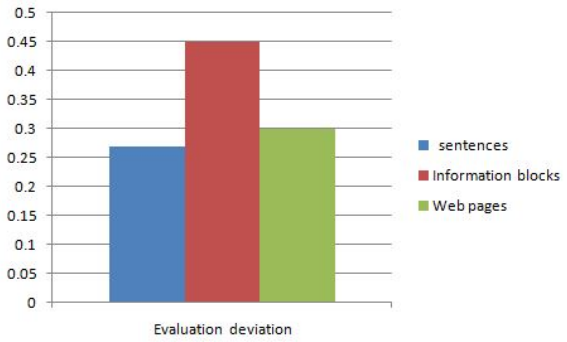

Figure 4. Comparison chart of three kinds of text

The progressive relationship and transition relationship between multiple sentences and other problems in the information Block are all likely to affect the tendency of the whole information block. So the evolution deviation of information block is bigger.

But in the web text we only consider the tendency of theme, so the difference between the results of manual evaluation is small.

VII. CONCLUSION

In this study, we try to find an effecetive method to achieve Tibetan text tendency. First according to the context and characteristics of Tibetan characteristics, combined with Hownet emotional release, we build the Tibetan word set. And then through the Tibetan word semantic similarity calculation, the dictionary is extended. Polarity of every word in the set is determined. We also proposed a method to obtain sentence tendency by the emotional value of emotional words in the sentence. And then through the cumulative value of the tendency value of sentences to get text orientation.

We select some blog from the websites and judge the tendency of them with our method. The results show that this method is effective.

VIII. ACKNOWLEDGEMENT

The work in this paper is supported by the National Natural Science Foundation of China project "Research on Basic Theory and Key Technology of Cross Language Social Public Opinion Analysis"(61331013).

REFERENCES

[1] Huang Xuan-Jing, Zhao jun. Chinese text sentiment orientation analysis. The first session of tendentious analysis of the national evaluation of Chinese, 2008

[2] Pang B, Lee L, Vaithyanathan S. Thumbs up Sentiment classification using machine learning techniques, Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg:Association for Computational Linguistics,2002:79-86.

[3] BoPang,Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL 2004, pp.271-278.

[4] Ni X ,Xue G,Ling X ,et al. Exploring in the Weblog space by detecting informative and affective articles. In Proc .of the16th International Conference on World Wide Web, 2007: 281-290.

[5] Cui H ,Mittal V ,Datar M .Comparative experiments on sentiment classification for online product reviews, In Proceeding of the 21th National conference on Artificial Intelligence (AAAI-06), Boston,USA,2006

[6] Ye Q, ShiW, LiY J. Sentiment classification for reviews: comparison between SVM and semantic approaches. In proceeding of The Fourth International Conference onMachine and Cybernetics. Guangzhou, 2005: 2341-2346.

[7] YeQ, ShiW, LiY J. Sentimentclassification formovie reviews in Chinese by improved semantic oriented approach. In Proceedings of the 39thHawaii InternationalConference on System Sciences, 2006: 53-60.

[8] Cai Jianping, Lin-Lin Wang, Lin Shiping Words and sentences polarity analysis based on machine learning, China Association for Artificial Intelligence the 12th National Conference Proceedings: Part One. Beijing: Beijing University of Posts and Telecommunications Press, 2007.

[9] Zhu Yan Lan, Min Jin, ZHOU Ya-Qian, etc. semantic orientation computing based on HowNet ,Chinese Information Technology, 2006, 20 (1): 14-20.

[10] Tumey Peter. Thumbs Up Or Thumbs Down Semantic Orientation Applied to Unsupervised Classification of Reviews. In proceeding of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 417-424.

[11] Hu Ming-qing, LIU Bing. Mining and summarizing customer reviews,Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2004: 168-177.

[12] Liu Bing,Hu Ming-qing,CHENG Jun-sheng. Opinion observer: analyzing and comparing opinions on the Web. In Proceeding of the 14th International Conference on World Wide Web. New York: ACM Press, 2005: 342-351.

[13] Kobayashi N,Inui K,Matsumoto Y,et al. Collecting evaluative expressions for opinion extraction. In Proceeding of the 1st International Joint Conference on Natural Language Processing. Berlin:Springer, 2005: 596-605.

[14] Ye qiang, Shi Wen, Li Yi-jun. Sentiment classification for movie reviews in Chinese by proved semantic oriented approach. In Proceeding of the 39th Annual Hawai International Conference on System Science, 2006.Washington DC: IEEE Computer Society,2006: 1-5.

[15] He ting-ting, Wen bin, Song le, etc. Research on word emotion recognition and opinion mining, the first session of the Chinese propensity analysis evaluation, Beijing, 2008:89-93.

[16] Quyou Li, Yu Hao, Xu Guowei, et al. Automatic segmentation of Web page information block [J]. Chinese Information Technology, 2003 (18): 4 - 13.

[17] Liu Weiping, Zhu Yanhui, Li Chunliang, et al. Research on building Chinese basic semantic lexicon. Journal of Computer Applications, 2009, 29 (10) :2875 -2877.

[18] Dao jieben, Cai rangcuo, Zhang tongling, research on Tibetan adverbs based on corpus, Northwest University for Nationalities (Natural Science), vol. 32, No.4, 2011 .12

[19] Yao Tianfang, Lou Decheng. Research on Chinese Semantic orientation discrimination, ICCC2007:The Seventh International Conference on Chinese Information Processing, Beijing: Electronic Industry Press, 2007:221-225.

[20] Qiu-yu Zhang, Peng Wang, Hui-juan Yang. Applications of Text Clustering Based on Semantic Body for Chinese Spam Filtering[J].Journal of computer, vol 7,No 11,2012:2612-2616.

[21] Vasileios Hatzivassiloglou, Kathleen R. Mckeown.Predicting the semantic orientation of adjectives: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics.Madrid,Spain: 1997:174-181.

[22] Turney P.Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, 2002:417-424.

[23] Turney P, Littman M. Measuring praise and cr iticism: inference of semantic orientation from association[J].ACM Transactionson Information Systems, 2003, 21(4) : 315-346

[24] Yucong Duan, Christophe Cruz, Christophe Nicolle. Identifying Objective True/False from Subjective Yes/No Semantic based on OWA and CWA [J].Journal of computer, vol 8, No 7,2013:1847-1852.

Xiaodong Yan was born in ChiFeng, China, in 1973. She received her Bachelor degree in electronic information area from Inner Mongolia University, Huhehot, China, in 1995, and her Master degree in electronic information from Peking University, Beijing, China in 1998 and Ph.D. degree in computer science from Beijing University of Posts and Telecommunication, Beijing, China, in 2006.

Since 2006, she is an Associate Professor in School of Information Engineering, Minzu University of China, Beijing, China. She has three books published: Minority language information processing overview (Beijing ,China, Minzu university Press,2012); new technology on searching multilingual network resources(Beijing ,China, Minzu university Press,2009) and grid computing(Beijing ,China, Minzu university Press,2007); She has more than 30 papers published and taken charge 6 projects on Natural Language Processing.

Xiaobing Zhao was born in Huhehot, China, in 1967. She received her Bachelor degree in computer science area from Inner Mongolia University, Huhehot, China, in 1988, and her Master degree in artificial intelligence from Korea Albatron schools, Hongseong, Korea in 2002 and Ph.D. degree in natural language processing from Beijing language University, Beijing, China, in 2007.

Since 2007, she is an Professor in School of Information Engineering, Minzu University of China, Beijing, China. She has more than 50 papers and 5 books published. And has taken charge 10 projects on Natural Language.

Professor Zhao has won the first prize of “Qian Weichang Chinese information processing science” and the third prize of “Inner Mongolia Science and technological progress Achievement Award “and so on.