International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.9 (2016), pp.385-394
http://dx.doi.org/10.14257/ijmue.2016.11.9.38

Research of Sentiment Classification for Tibetan Texts by Supervised Learning

Lirong Qiu and Zhen Zhang

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

Abstract

Increasing number of subjective text appears on the internet which contains a lot of information. In this paper, we study how to apply supervised learning techniques to solve sentiment classification problems. Using the Tibetan news as data, we find that standard supervised learning techniques definitively outperform human-produced baselines. Moreover, we find that selecting the words with words with polarity as feature, the special syntactic structure such as exclamation sentence pattern, etc. as feature can improve the performance of sentiment classification. Conclusively, the research of sentiment analysis is a more challenging problem.

Keywords*: Tibetan information processing; sentiment analysis; text classification; feature selection*

1. Introduction

Text classification has been one of the key tools to automatically handle and organize text information for decades. In recent years, with more and more subjective information appearing on the internet, sentiment classification [1], as a special case of text classification for subjective texts, is becoming a hotspot in many research fields, including natural language processing (NLP), data mining (DM) and information retrieval (IR).

The rapid development and widespread use of the online shopping, blogs, forums and social networking platforms and the subjective statements on Internet are getting larger. These texts contain a lot of valuable information, for example, the users' evaluation and ideas towards social phenomena and products, it not only produces great influence on people's consumption habits and the social public opinion, but also very meaningful for social stability, development of enterprises and even individuals. Therefore, the urgent demand of mining and expressing the deeper semantic information from the mass of unstructured data is showing more significance. The sentiment classification has started a few years, which study how to classify the subjective contexts and analyze its polarity is positive or negative. There are many important practical applications:

(1)Filtration system: It is used to filter adverse speech for the government and commercial organization and classify the text based on the subjective context.

(2)Recommended system: Pick up the recommended goods or services and push to users according to analysis and classify online feedback from the users. For example, analysis the movie or book reviews, to decide whether to make a recommendation to the user.

(3)Prediction of stock market: According to the emotional tendencies of Stock analysts and the links between the yield of the market and abnormal postings, to study the correlation between the stock analysts emotion and excess returns and predict the stock market movements.

The paper studies how to use supervised learning approach to classify the subjectivity news. The emotion-based text classification has the similar with the topic-based text

ISSN: 1975-0080 IJMUE

classification and the difference is the selection of feature. Topic-based text classification is more emphasis on the subject words while the emotion-based text classification prefers to the subjective words, such as “crisis”, “boycott”, “excellent” *etc.*

2. Related Work

The earliest text classification is based on the topic which is to determine a category for the documents in accordance with pre-defined subject categories and classified by subject characteristics. The topic-based text classification algorithm has been very mature and the currently main research of the automatic text categorization based on statistical methods and machine learning. In recent years, with the development and needs of the society, the emotional-based text classification also begins to be gradually concerned. The sentiment analysis mainly contains the semantic tendency recognition of the words, the text sentiment classification, extraction viewpoint and the subjective analysis.

According to size of the classification text, the sentiment classification can be divided into phrases level and sentence level and text level. The text level is the focus of the paper and there are three classification methods: the unsupervised learning and the supervised learning. According to proportion of the marked sample in the training set, text-level sentiment classification can be classified the supervised learning and the unsupervised learning (the rule-based method).

2.1. The Supervised Learning Method

The supervised learning trains classification model through a large number of labeled samples and classify the texts with the trained model. PangBo [2] applied supervised learning method to solve text sentiment classification of film comment text for the first time in 2002, and compared performance of a variety of features (unigram, bigram, part-of-speech) and weights (boolean, tf) applied to the classification. PangBo [3] transformed the subjective sentences classification into the minimum cut of the figure and achieved a cut-based classifier applied to the emotion recognition. Abbasi [4] applied the information gain to select the features that contributed to the sentiment classification. 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. Yuan Bin [8] combined Tibetan semantic features and K-means method for Tibetan Weibo sentiment analysis.

2.2. The Unsupervised Learning Method

Compared with the supervised learning sentiment analysis, the research of the unsupervised learning or rule-based method is not many. In addition to Turney [9] (2002), Zhu Yanlan [10] used HowNet to calculate the emotional tendency of Chinese words semantic meaning. Lou Decheng [11] used syntactic structure and dependence relationship to analyze the emotion of Chinese sentence. Hiroshi [12] realize Japanese phrase-level sentiment analysis through a rule-based machine translator. On the basis of SO-PMI algorithm proposed by Turney, Zagibalov [13] analyzed 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. Zhang Jun [14] take the based on sentiment dictionary method to analysis sentiment of the Tibetan microblogging and

achieved good results.

3. Classification Algorithm

Our target in this work is to examine whether it suffices to treat sentiment classification simply as a special case of topic-based classification (with the two "topics" being positive sentiment and negative sentiment), or whether special sentiment classification methods need to be developed. We experimented with two standard algorithms: Naive Bayes classification and maximum entropy classification. The philosophies behind these three algorithms are quite different, but each has been shown to be effective in previous text categorization studies.

The sentiment classification of news is a binary classification problem and the target categories are positive and negative. We adopt the machine learning which is used to topic-based text classification with better performance, to realize the sentiment classification of the text. In vector model of text, the document features can be expressed as $\{f_1, f_2, f_3, \ldots, f_m\}$, and there are **m** features totally. The document **d** is expressed as $\{n_1, n_2, n_3, \ldots, n_m\}$, and n_i represents the weight of the f_i. The number of the training set that belong to the category c_j which is represented as N_j, and **N** represents the number of the training set, $n_{i,j}$ represents the number of times that the feature f_i appears in the category c_j.

3.1. Naïve Bayes

Naïve Bayes is the probabilistic classifier which uses the prior probability of the category and the conditional probability of the feature relative to the category to calculate the probability of the unmark document belonged to a category. Suppose that document features distribute independently of each other, Naïve Bayes can be expressed in mathematical form as follows.

$$P(c_j|d) = \frac{P(c_j)P(d|c_j)}{P(d)} = \frac{P(c_j)\prod_{i=1}^{m} P(f_i|c_j)}{P(d)} \quad (1)$$

For different categories, the denominator is constant, so choice the category that can maximize numerator, which is the target category of the unmark documents. Obtain valuations of $P(c_j)$ and $P(f_i|c_j)$ through the training and learning of samples:

$$\hat{P}(c_j) = \frac{N_j}{N} \quad (2)$$

$$\hat{P}(f_i|c_j) = \frac{1+n_{i,j}}{m+\sum_{k=1}^{m} n_{k,j}} \quad (3)$$

Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes-based text categorization still tends to perform surprisingly well (Lewis, 1998) [15]; indeed, Domingos and Pazzani (1997) show that Naive Bayes is optimal for certain problem classes with highly dependent features [16]. The most fundamental characteristic of the Naïve Bayes is occurrence independent of the documents words. The probabilistic method is uncomplicated and effective classification method. The category of the document depends on the category of the maximum characteristics probability. So, Naïve Bayes classifier can be well applied to classification.

3.2. Maximum Entropy

The basic idea of Maximum Entropy is that seek the most balance model under the

condition of meeting the current system requirements. It choices the probability distribution of making entropy maximization as the correct probability distribution under the constraints and select the data item related to the classification as a series of features which is represented by binary function in most cases.

For the sentiment classification of text, we choice the pair of feature word and category ($f_i - c$) as one feature and decide the value of feature is binary or frequency depend on requirements, which is better adapt to the text classification. The feature function form as follows:

$$F_{i,c}(d,x) = \begin{cases} n_i, if\ n_i > 0\ and\ x = c \\ 0, otherwise \end{cases} \quad (4)$$

For the feature function $F_{i,c}$, expectation of the relative empirical probability distribution $\hat{P}(f_i|c)$ is consistent with the expectation of the relative model $P(c|f_i)$, so get the constraint function of the probability distribution:

$$E_P F_{i,c} = E_{\hat{P}} F_{i,c} \quad (5)$$

Introduction of the lagrangian multiplier and get the equation:

$$P^*(c|d) = \frac{1}{\pi(d)} e^{\left(\sum_{i=1}^{m} \lambda_{i,c} F_{i,c}(d,c)\right)} \quad (6)$$

Here, $\pi(d)$ is the normalization factor and the calculation equation shown as following:

$$\pi(d) = \sum_{c \in C} e^{\left(\sum_{i=1}^{m} \lambda_{i,c} F_{i,c}(d,c)\right)} \quad (7)$$

The $\lambda_{i,c}$ is the parameter of feature function which is trained from the training set. The value of $\lambda_{i,c}$ is calculated by the IIS [17] algorithm. For the sentiment classification of text, the feature is very sparse and most of feature function value is 0, which is need to use smoothing techniques processing. There is no one specific smoothing techniques for Maximum Entropy Model. For text sentiment classification, we take the Absolute-Discounting as the smoothing technique, which discounts the observed event by minus a fixed value d and accumulates the values to divide equally to the unobserved events. In the processing of discounting the term frequency, because of frequency feature function, there is no need to consider to maintain sum of the probabilities is 1, so, here, assign a fixed value ad for the feature whose term frequency is 0. After the processing of Absolute-Discounting, the feature function (4) based on term frequency becomes the following form:

$$F_{i,c}(d,x) = \begin{cases} n_i, if\ n_i > 0\ and\ x = c \\ ad, otherwise \end{cases} \quad (8)$$

4. The Method of Generating Tibetan Text Feature

4.1. The Classification Model

The basic model of text categorization shown in Figure 1, the classification consists of the training process and the testing process. In training process, firstly, generate features of the training text and obtain the feature set; then, selection algorithm to extract the optimal subset of features from the text feature ensemble, and the optimal subset is determined by the evaluation algorithm which classifies the text feature subsets that represented the text and evaluates the performance of the classifier. In the testing process, it handles that the optimal feature subset represents a test document and then classify the

feature subset by classifier and obtain the category of the test document.

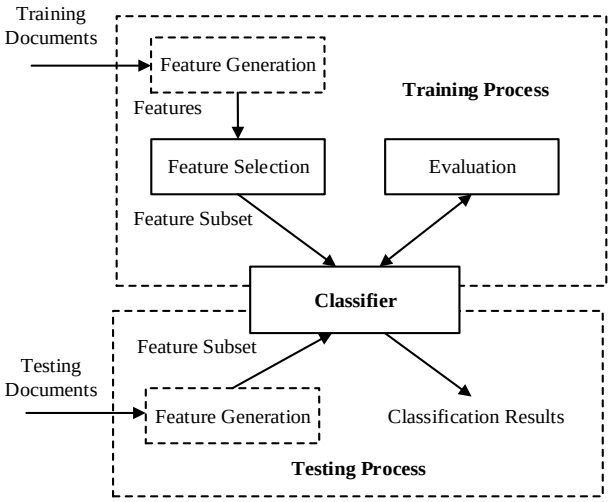
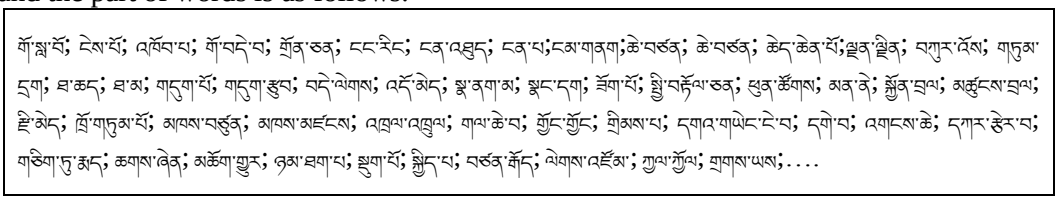


Figure 1. Mode of Text Categorization

3.3. The Experimental Data

The resources of Tibetan corpus are very limited and there is almost no Tibetan Forum on line. For our experiments, we chose to work with Tibetan news crawled from the People Tibetan plate. There are a large number of neutral articles on news websites and directly uses these texts for the experiment, experimental effect will reduce. Therefore, we need to deal with experimental data for the further selection.

Step 1: Pick out 100 the strong Emotional words (including some adjectives and slightly adverbs) as the basic emotional lexicon that Tibetan news is most likely to occur and the part of words is as follows.



Step 2: Select the target new text by matching emotional lexicon, if the text contains more than 10 percent of the word lexicon, we believe that the text has emotional tendencies, and meets the experiment requirements.

Step 3: The text Processing: remove the sentence contained the non-Tibetan characters in the text and take word segmentation and Part of Speech processing of the sentences, which is achieved by Yangjin Tibetan segmentation system [18] that developed by Institute of Machine Translation, Northwest Nationalities University.

Step 4: Training corpus annotation: For corpus annotation, we take artificial annotated method.

Emotional Labeling is a binary classification criteria, which do not need to do much work and processing of the labeling is not easy to make mistakes, so it can quickly and efficiently accomplished by artificial means.

By processing the above steps, obtain ideal experimental data and use for experiments.

3.4. The Experiments

As for the selection of the feature items, here, it choices the term frequency and binary

as the text feature weights and the performance of term frequency precedes the binary in the topic-based text classification [19]. To analyze the effect of the different weight calculation method on the sentiment classification text, we conduct experiment (I) to compare performances of the above methods.

Consider that the appearance of negative words will impact on the emotion of text, so we conduct experiment (II) to process the negative words. Here, negative word itself is a neutral word and when an object with the polarity is modified by the negative word, the semantic orientation is reversed. Tibetan grammar is different from English and Chinese, and the negative word immediately follow the modified objects, for example, གཏན་རྒྱལ་གང་ཞེག་སྐྱག་གྱང་ང་ཁྱེད་དང་ཐ་བྲལ་གྱི་མེན། (no matter what happens I will not leave you), among the words, ཐ་བྲལ (“leave”) is the modified object and གྱི་མེན (“not”) is the negative word. In the paper, we combine the modified object and negative word as a new feature and do the appropriate experiment.

According to Tibetan language features, in Tibetan sentiment sentences, emotional performance is expressed by nouns, verbs, adjectives, interjections, modal particles, specific sentence, and the degree adverbs play an important role in document emotional tendencies, which also extracted as a feature. Table. 1 lists common part of speech with the emotional tendencies. To analyze the differences of part of speech act on emotional text, we conduct experiment (III) to verify the performance of different parts of speech. Using the chi-square statistic method to analysis the independence between variables. The chi-square statistic method can measure the independence of feature item t and category c in the feature selection. If feature item t is independent of the category c, it indicates that the feature item does not play a characterized role for the category c. On the contrary, it can be used to select the feature items that play the role characterized role for classification.

The formula is as follows.

$$CHI(t) = \sum P(C_i)\chi^2(t, C_i) = \sum P(C_i)\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (9)$$

Here, the A represents the frequency of the feature item t and the document category c_i occurring simultaneously; B represents frequency of the feature item t but the document category c_i not being occurring simultaneously; C represents the frequency of the document category c_i but the feature item t not being occurring; D represents the frequency of the document category c_i and the feature item t not being occurring. From the statistical point of view, if $CHI(t)$ is equal to 0, the feature item t does not contain any information of the category c_i. The more relevant between feature item t and the category c_i, the larger value of $CHI(t)$ and the more information which is relevant to the category c_i. Therefore, we designed the experiment (IV) that is based on N-POS model feature selection algorithm, and analysis the effect that the combination of different parts of speech for text sentiment classification. In the subjective and objective corpus, we counted all 2-POS items and 3-POS items that can express the text emotional tendencies and the statistical results are shown in Table 2 and Table 3

Table 1. Common Semantic Tendency Part of Speech and Some Examples

POS	Tag	Part of words
adjective	A	མཛེས་(beautiful); ཐ་ཤལ་(ugly); ཕང་ཕང་(healthy)
noun	N	ལེགས་བྱས་(good deed);ས་འཐུལ་(earthquake); མེ་སྐྱོན་(fire disaster)
verb	V	སྒེལ་བ་(develop); རྩོམ་(rob); འཇོམས་པ་(plunder)
adverb	D	ཚ་དྲག་(very); བཟའ་བར་(secretly); ཤོས་(most)
particle	U	ཀྱི་དུད་(sigh); ཀྱེ་མ་(alas); ཀྱེ་མ་རྟོ་(ah)

Table 2. 2-Pos Items Mode Statistics

First word	Last word	Subjective	Objective	χ^2 Statistics
Adverb	adjective	836	12	114.2442
Adverb	verb	1918	50	202.8557
Adjective	verb	228	16	76.3462
Verb	adverb	684	30	74.5646
Verb	adjective	352	14	67.9839

Table 3. 3-Pos Items Mode Statistics

First word	Middle word	Last word	Subjective	Objective	χ^2 Statistics
Adverb	adjective	auxiliary	66	6	49.3797
Verb	adverb	adjective	58	0	45.9779
Adverb	verb	Verb	94	4	40.4807
Verb	adverb	Verb	100	6	35.6990
adjective	adjective	auxiliary	20	0	30.6397

As can be seen from Table 3, the number of the subjective corpus that expressed by 3-pos items is less than 2-pos items while the tendency that expressed by 3-pos items is more accurate than 2-pos items.

4. Experimental Results and Analysis

The experimental data is crawled from the People Tibetan plate (http://tibet.people.com.cn/) and after the experiment data processing, we select 5000 new texts which is in a 1: 1 ratio randomly divided into the training set and the testing set. Conduct 50 tests on the testing set and compare with the average accuracy rate. The results are shown in Table 4, 5 and 6.

Table 5. The Accuracy of Different Weight Calculation Method

No	weight	accuracy	
		NB	ME
1	TF	76.67	79.07
2	Binary	77.12	80.09

The weight calculation method of binary is superior to the term frequency (TF) as can be seen from Table 2, which is significantly different from the Topic-based text classification. The main reason is that emotional words seldom repeated in the document

and once the word appears that can determine the emotional tendencies of sentence or document, which is regardless of term frequency.

Table 5. The Accuracy of Processing Negative Word Classification

No	Weight	accuracy	
		NB	ME
3	TF	77.32	79.02
4	Binary	78.56	80.20

The process of negative words can help us to improve appropriately sentiment classification accuracy. Tibetan vocabulary is very rich, such as synonym, antonym. In the new text, negative words rarely appear which is replaced with antonyms. The experimental effect is not obvious.

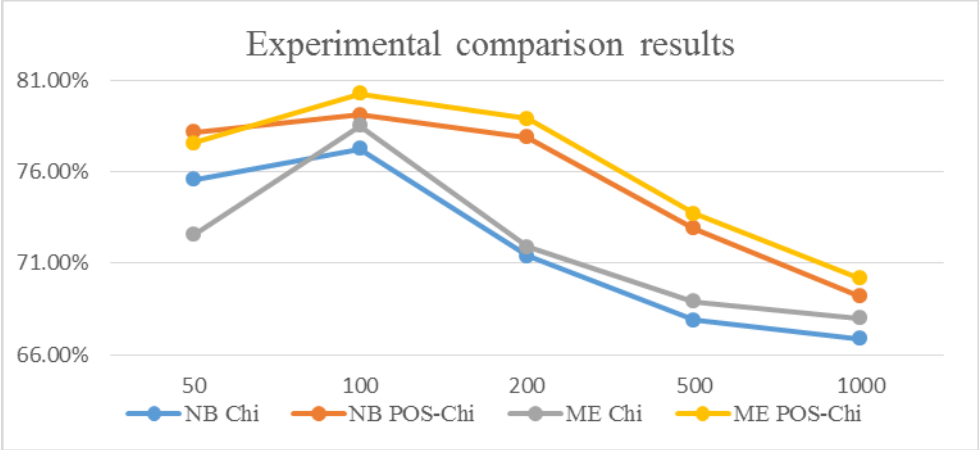


Table 6. The Accuracy of Processing Different Parts of Speech Classification

No	Negative	POS	Accuracy	
			NB	ME
5	No	A	70.32	69.48
6	Yes	A	72.24	72.39
7	Yes	A,N	74.60	76.12
8	Yes	A,N,V,D,U	79.11	80.25

Figure 2. Experimental Results of Different Feature Items and Algorithms

As can be seen from the Figure2, when the characteristic dimension is 100, the classification results can be achieved best. Compare experiment results 5 and 6, it can be seen that, to a certain extent, negative words can effectively improve the classification effect under the case of a few feature items. Some word does not have the emotional tendencies, but show the sentiment orientation or reverse emotional polarity when appears with some words together, such as “ཉེན་ཁ་དག་འཕྲོད (get out of danger)”, “འཕྲོད (change)” is a neutral word, and “ཉེན་ཁ(danger)” is negative word, but the entire phrase has shown positive emotional tendencies. The negative reverses the semantic orientation of the

modified words, such as, "ཡང་དག་མིན་ད (unreal)", among them "ཡང་དག (real)" is positive word, and "མིན་ད (not)" is a negative word. Because adjectives, nouns, verbs, adverbs are the most common words with semantic tendencies, the using of them decide the sentiment orientation of document. From the experiment result 6, 7 and 8, it can find that the more emotional feature items to be considered, the better classification effect.

5. Conclusions

In the paper, we take the Naive Bayes and Maximum Entropy to classify the Tibetan Emotional text and the Maximum Entropy method is superior to Naive Bayes in most cases. Using binary as the feature weight is better than term frequency. During the experiment processing, it found that the sentiment orientation words play a decisive role in emotional classification and the effective selection of sentiment feature is contribute to emotional classification. As previously analyzed, some word does not have the emotional tendencies, but show the sentiment orientation when appears with some words together, so that we design a method for generating a bigram item to improve the classification performance, and also improve efficiency by identifying effectively objective sentence and removing them from the articles. In future work, we will explore even richer linguistic analysis, for example, parsing, semantic analysis and topic modeling.

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. Liu, "Sentiment analysis: a multifaceted problem", IEEE Intelligent System, vol. 25, **(2010)**, pp. 76–80.
[2] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques", In: Proceeding of the 2002 Conference on Empirical Methods in Natural Language Processing, **(2002)**, pp. 79-86.
[3] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of ACL-04, pp. 217-278.
[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", Proceedings of KDD, **(2009)**, pp. 1275-1283.
[7] M. Taboada and J. Brooke, "Manfred Stede. Genre-Based Paragraph Classification for Sentiment Analysis", Proceedings of SIGDIAL, **(2009)**, pp. 62-70.
[8] Y. Bin, J. Tao and Y. H. Zhi, "Tibetan Microblogging Sentiment Analysis Based on The Semantic Space", Application Research of Computers, vol. 2016, no. 3.
[9] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of ACL-02, pp. 417-424.
[10] Z. Y. Lan and M. J. Zhou, "Semantic Orientation Computing Based on How Net", Journal of Chinese information processing, vol. 20, no. 1, **(2006)**, pp. 14-20.
[11] 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.
[12] K. Hiroshi, N. Tetsuya and W. Hideo, "Deep Sentiment Analysis Using Machine Translation Technology", Proceedings of COLING, **(2004)**.
[13] T. Zagibalov and J. Carroll, "Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text", Proceedings of COLING, **(2008)**, pp. 1073-1080.
[14] Z. Jun and L. Y. Xing, "Tibetan Microblogging Sentiment Analysis Based on The Sentiment Dictionary", Silicon Valley, vol. 2014, no. 20.
[15] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval", In Proceeding of the European Conference on Machine Learning (ECML), Invited talk, **(1998)**, pp. 4–15.
[16] P. Domingos and M. J. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss",

Machine Learning, vol. 29, no. 2-3, **(1997)**, pp. 103-130.
[17] R. Adwait, "A simple introduction to maximum entropy models for natural language processing. Institute for Research in Cognitive Science, University of Pennsylvania. Tech. Rep. 97-08, **(1997)**.
[18] S. X. Dong and L. Y. Jun, "A Tibetan Segmentation System—YangJin", Journal of Chinese Information Processing, vol. 2011, no. 4.
[19] M. Lan and S. Y. Sung, "A Comparative Study on term Weight Schemes for Text Categorization", International Joint Conference on Neural Networks, **(2005)**.

