Chinese sentiment analysis for commodity price level fluctuation news comments

Yan Zhao[1*], Suyu Dong[1], Jing Yang[2]

[1]*College of Management, Inner Mongolia University of Technology, Huhhot, China*

[2]*College of Mechanics, Inner Mongolia University of Technology, Huhhot, China*

Received 1 March 2014, www.cmnt.lv

Abstract

With the rapid development of the Internet technology and news media, people pay more attention on news especially about commodity price fluctuation. Hence, more and more Chinese news comments about commodity price fluctuation appear on Internet. These comments contain all kinds of sentiment. Analysing the sentiment of these comments will make government know more about Netizen emotion on this information and enhance efficiency of management, which has important practical significance. In this paper, we adopt three supervised learning methods (naive Bayes, maximum entropy and support vector machines) to automatically classify user comments as two classes (positive and negative). Through a lot of experiments, we found that machine learning techniques perform quite well in the domain of commodity price fluctuation news comments sentiment classification. Meanwhile, the effects of the feature representations and dimensions for the classification of the three machine learning techniques are analysed and discussed in detail. Experimental results show that maximum entropy classifier is best overall. Frequency is a better method of feature representation, which can use fewer features to get better result.

Keywords: sentiment classification, online reviews, supervised machine learning algorithm

1 Introduction

Nowadays, with the rapid development of the news media, Internet becomes an important part of persons' daily life. The news can be spread with high speed through Internet, which adds people's awareness of paying close attention to online news. Hence, most persons are more likely to comment some news, especially to the news closely linked with daily life, such as news about commodity price fluctuation. Most Netizen comments contain opinion or sentiment, which can impact persons, society and government. As the managers of society, government needs to know the public opinion in time. However, it is time-consuming to browse and analyse all the comments artificially. Hence, an efficient and automatic method of analysis and statistics is necessary.

The technology of sentiment classification aims at analysing texts' opinion automatically. Sentiment classification is widely used in many fields, such as consumption of product, service, social events, vote and so on. Now, in the domain of sentiment classification, most researchers focused on product and service field, few scholars research the classification of new comments. However, news comments contain a lot of public sentiment information about government policies, especially the policies about commodity price fluctuation, which directly impact the person daily life. What's more, people pay more attention to the policies about commodity price fluctuation. Hence, the analysis and research about these news comments is very necessary.

In this paper, we adopt the methods of machine learning to analyse the sentiment of comments about commodity price fluctuation. We select the three popular machine learning algorithms (naive Bayes, maximum entropy and support vector machines) to classify the sentiment of comments. We also experiment with three feature representation methods (presence, TF, TF-IDF) in the conditions of different feature dimensions. Hence, this paper will analyse and discuss the following problems:

1) Which is the best classifier among SVM, naive Bayes and maximum entropy regarding sentiment classification of commodity price fluctuation news comments?

2) Which method of feature representation for classifiers is the best method regarding sentiment classification of commodity price fluctuation news comments?

3) How the feature dimensions affect the result of classification?

2 Previous Work

With the development of society, Internet has become indispensable part of people's life. Persons pay more and more attention to online news, especially the news closely linked with persons' life. As an important aspect of product and consumption, commodity price fluctuation

* ***Corresponding author*** e-mail: yanzhaosky@126.com

has become persons' focus of attention. Nowadays, the clicks of news about commodity price fluctuation rise incessantly, more and more Netizens comment online. In this situation, the large numbers of comments coming from Netizens have had huge impact on society life and public opinion. For the government, the technology of text classification can improve the efficiency of grasping the Netizen comment information.

Sentiment classification aiming at classifying texts according to positive or negative emotion. Sentiment classification has become the research hotpots now, which is a branch of the natural language processing. What's more, the sentiment classification also has large impact on other domains, such as management, sociology, economics and so on. For example, the technology of sentiment classification can help consumers know the information of comments coming from other consumers, enterprise also can grasp the opinion of consumers about the commodities. In addition, government also can grasp the opinion of Netizens as well as the proposals of public. Some researches have applied sentiment classification to some practice fields, such as product, service, financial and so on [1-9], This paper will research the sentiment classification on the news comments of commodity price fluctuation.

In the domain of sentiment analysis, sentiment classification is the most wide research aspect [10-13]. Existing methods of sentiment classification on learning machine can be divided into two classes, which are supervised and unsupervised classification methods. Sentiment classification is one of text classification researches. Hence, existing methods for text classification can be useful for sentiment classification. As a kind of short texts, news comments also can be classified through the methods of traditional text classification. Positive and negative emotions expressed by the emotional words are one of the most important indicators in the problems of sentiment classification. Hence, some researchers used emotional words to classification with unsupervised learning methods [14-17]. [14] used sentiment words, sentiment phrases and fixed syntactic patterns which are used to express sentiment to perform sentiment classification. [15-17] used emotional direction and strength of emotion word, phrase emotional dictionary to sentiment classification, which adopted emotion strength and negative conversion to calculate emotional value to perform classification.

Compared with unsupervised learning methods, supervised machine learning methods are more often used on sentiment classification. Most popular machine learning methods are naive Bayes, maximum entropy and support vector machines. [1] compared naive Bayes, maximum entropy and support vector machines on movies reviews sentiment classification, whose result showed that the feature representation method of presence is better than TF, naive Bayes is better than SVM with TF, SVM with presence is better than naive Bayes, maximum entropy is worst among the three classifiers. [18] compared SVM, naive Bayes, N-gram semantic model and found that SVM and N-gram semantic model are better than naive Bayes on classification of travel reviews. [19] proved that SVM is not better than naive Bayes for literature reviews classification at all times. [20] showed that naive Bayes is a better classifier than SVM. The examples listed above show that no machine learning algorithm can always maintain the best effect for all kind of texts. Hence, It is necessary to compare naive Bayes, maximum entropy and SVM to find the best methods in commodity price fluctuation news comments classification. In addition, other factors such as feature representation method and feature dimensions are also worth researching.

3 Methodologies

Figure 1 shows the fundamental steps of text classification using supervised learning machine methods. As is known from fundamental theory, the supervised learning machine methods can train labelled texts to get models, which can be used for predicting new unlabelled texts.

The following part of this section will introduce the theory of every section, including feature selection methods, feature representation methods, and classifiers. In this paper, each document is represented as a vector with feature weights. Let $\{t_1, t_2, ..., t_m\}$ be a predefined set of m features that can appear in a document. Let n_i be the number of times that t_i occurs in document d. Let w_i is the feature weight in a document. Each document d is represented by the document vector $d = \{w_1, w_2, ..., w_m\}$.

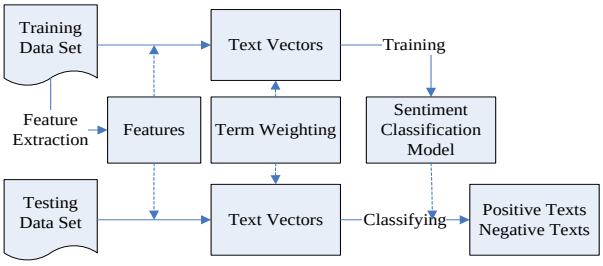

FIGURE 1 Fundamental theory of text classification

3.1 FEATURE SELECTION

Large numbers of features will be produced through feature identification. If all the features are used for classifier, the result and efficiency of classification will be reduced. Hence, feature selection is essential for classification. This paper adopts document frequency feature selection method, whose another name is DF. Generally, the DF firstly counts the number of every feature (DF value) appearing in all texts (comments), and then get the proper features according to the DF value. In our experiment, all the features are sorted according to the DF value. We select top n features to do experiments, The value of n is from 50 to 2950. The 30 groups of features will be selected, the distance of numbers is 100 among each group.

3.2 FEATURE REPRESENTATION

Feature weighting denotes the importance of a feature of a text, namely the distinguish ability of a feature of the text. Feature weight is calculated through the statistical information of text. This paper will compare three different calculation methods of feature weight for classification, they are Boolean, frequency and TF-IDF.

3.2.1 Presence representation

Boolean is based on the feature whether or not appears in the text. When the feature appears in the text, the value is 1, otherwise the value is 0. This method is replaced by other more accurate methods because the Boolean value cannot reflect the importance of feature in text. However, this method can also obtain a good effect under some circumstances. For instance, [1] performed the sentiment classification of movie reviews, which showed that SVM with unigrams model combining Boolean feature representation is better than the other methods of feature representation in the same circumstances.

3.2.2 TF

The method of TF uses the times of feature appearance in the text to represent the text. When we use frequency as the calculation of feature weight, the distinguish ability of low-frequency features will be ignored. However, some low-frequency features may have a greater ability to distinguish text than high-frequency features. [21] used naive Bayes with TF feature representation method performed the standard topic-based categorization and the accuracy is highest. [19] found that frequency is the better feature representation method for SVM and naive Bayes in the novel text sentiment classification experiment.

3.2.3 TF-IDF

TF-IDF is the most widely used feature weight calculation method for the text classification. It is based on the idea: if one feature has high-frequency, and rarely appears in other text, then the feature has a good ability to distinguish. Although it is ideas and structure of statistics are very simple, but its performance is very good. The $TF-IDF$ value of a certain feature is calculated by the following equation:

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_i}, \tag{1}$$

where w_{ij} indicates the weight of feature t_i in document d_j. tf_{ij} indicates the frequency of feature t_i in document d_j. n_i indicates the number of document which contains feature t_i. N is the number of all documents.

3.3 CLASSIFIERS

3.3.1 Naive Bayes classifier

Naive Bayes classifier is a kind of simple classifier, but widely used in text classification. According to the Bayes formula, the probability of document d belongs to C_i is calculated by the following equation:

$$P(C_i|d) = \frac{P(d|C_i) * P(C_i)}{P(d)}, \tag{2}$$

where $P(C_i)$ indicates the probability of a document belonging to C_i. In this paper, we used Naive Bayes classifier with weight. The equation is following:

$$P_{NB}(C_i|d) = \frac{P(C_i)\left(\prod_{t_i \in V} P(t_i|C_i)^{W(t_i,d)}\right)}{\sum_j \left[P(C_j)\prod_{t_i \in V} P(t_i|C_j)^{W(t_i,d)}\right]}, \tag{3}$$

where feature t_i is independent of document d, $W(t_i,d)$ indicates the weights of feature t_i in document d. $P(t_i|C_i)$ indicates the Laplacean probability estimation value of conditional probability of documents belonging to C_i if it contains feature t_i. $P(t_i|C_i)$ is calculated by the following equation:

$$P(t_i|C_i) = \frac{1 + W(t_i, C_i)}{|V| + \sum_j W(t_j, C_i)}, \tag{4}$$

where $W(t_i, C_i)$ indicates the number of documents containing features t_i and belonging to C_i. $|V|$ is the size of $\{t_1, t_2, ..., t_m\}$, which are all features coming from all documents.

Naive Bayes classifier is based on the assumption of independence conditions, using the joint probability between features and categories to estimate the probability of categories given a document. Although it's assumption conditions is very restrictive and difficult to meet in real-world, it still performed well in text classification [20,22]. [23] showed that naive Bayes can well complete two opposite case data classification, completely independent features classification or functionally dependent features classification.

3.3.2 Maximum entropy classifier

Maximum entropy classifier (ME) is based on maximum entropy model, [24] was the first application of maximum entropy models in the natural language processing; [25] improved maximum entropy model. [26] found that ME is better classifier than Naive Bayes classifier on text classification. Its basic idea is that it does not make any

hypothesis and remain maximum entropy for the unknown information, this is an advantage for maximum entropy compared with Naive Bayes. Maximum entropy model must satisfy the constraint of known information and the principle of maximum entropy. Hence maximum entropy model is got through solving a optimization problem with constraints. The classical algorithm to solve this problem is lagrange multiplier method. In this paper, we give the conclusion directly. The result is following:

$$p^*(C_i|t_i)=\frac{1}{\sum_{C_i}\exp\left(\sum_i \lambda_i f(t_i,C_i)\right)}\exp\left(\sum_i \lambda_i f(t_i,C_i)\right), \quad (5)$$

where $p*$ indicates a predictive model for classification; V indicates the feature vectors; C_i indicates the type which the document belongs to. λ_i indicates the feature weight of feature vectors containing many feature t_i. $f(t_i,C_i)$ is an indicator function.

3.3.3 SVM

Support vector machine (SVM) is generally considered as the best classifier for traditional text classification [27], it is usually better than naive Bayes and maximum entropy. Naive Bayes and maximum entropy are based on probability model, support vector machine (SVM) classifier is got by solving the optimal hyperplane represented by vector $\vec{W}$. Hyperplane is shown in Figure 2, this hyperplane is used to accomplish classification which can ensure maximum separation between a certain amount of data from the training set and hyperplane. Solving the maximum margin hyperplane eventually is converted into solving a convex quadratic programming problem.

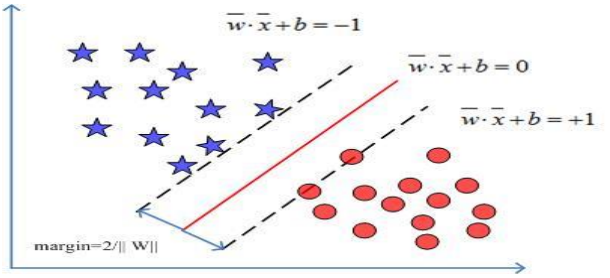

FIGURE 2 The optimal hyperplane

Generally, it translates the above problem into the constrained optimization problem of dual variables through Lagrange Duality. The solution can be written as:

$$\vec{W}=\sum_{i=1}^{n}\alpha_i C_i \vec{d}_i, \quad (6)$$

C_i is the correct category for document $\vec{d}_i$. α_i are support vector and greater than zero.

Moreover, kernel function can be used for linear inseparable problems for SVM to convert low dimensional space nonlinear problem to high dimension space linear problem. Mapping of kernel function can be a good control of the computational complexity of nonlinear expansion and can avoid the curse of dimensionality. There are many kernel functions: linear kernel, Gaussian kernel function, radial basis function and so on. In this paper, we used linear kernel function and optimize the parameter of SVM model, which will be used for following experiments.

4 Experiments

4.1 EXPERIMENT DATA

This paper research the sentiment classification of online comments to news of commodity price fluctuation, therefore, we created a corpus by retrieving reviews of news of commodity price fluctuation from each big Chinese news website (URL: http://news.sina.com.cn; http://news.sohu.com; http://news.qq.com; http://news.163.com; http://www.people.com.cn). The data we used for experiment were downloaded by a crawler and randomly selected 3566 comments from all comments.

In this paper, we focus on classifying comments as positive or negative. However, there is no label about the polarity of sentiment for comments. Thus, three students were trained to annotate these comments. In the whole process of the annotation, non-commodity-price-fluctuation news comments were excluded before annotating the polarity of sentiment. Comments were annotated polarity according to the unified label; the values were 0, 1, 2. Thereinto, 0 represents the negative comments, 1 represents the positive comments, 2 represents the comments which its polarity cannot be judged. We found that there was inconsistent between students when they annotated comments. We weeded out the comments if:

1) it is annotated differently by three students;
2) two students' judgment annotates it with 2.

Finally, 1500 negative comments and 1500 positive comments were randomly chosen to establish the corpus.

4.2 EVALUATION METHOD

In this paper, the results of sentiment classification are evaluated by three indexes that are frequently used in text classification: Accuracy, Precision and Recall. Accuracy is used to justify the overall performance of sentiment classification. Precision and recall are used to evaluate the performance the negative and positive classification. These indexes can be calculated according to Table 1.

TABLE 1 Results of experiments

	Classified positive comments	Classified negative comments
labelled positive comments	a	b
labelled negative comments	c	d

The calculation equations are the following, respectively:

$$\text{Accuracy}=\frac{a+d}{a+b+c+d}, \tag{7}$$

$$\text{Precision}(pos)=\frac{a}{a+b}, \tag{8}$$

$$\text{Precision}(neg)=\frac{d}{c+d}, \tag{9}$$

$$\text{Recall}(pos)=\frac{a}{a+c}, \tag{10}$$

$$\text{Recall}(neg)=\frac{d}{b+d}. \tag{11}$$

5 Experiment result and discussion

We adopt 3-fold cross-validation to do experiment. We adopt our own implementation for text pre-processing, NLPIR toolkit is used for Chinese text segmentation, McCallum's Mallet toolkit [28] implementation of naive Bayes classifier and maximum entropy classifier and Chang's LIBSVM [29] implementation of a Support Vector Machine classifier are used for classification. We ran each classifier with various feature representations and different number of features to experiment.

5.1 EXPERIMENT RESULTS OF NB CLASSIFIER

The performances of sentiment classification of NB with various feature representations and different feature sizes are showed in Figures 3-5.

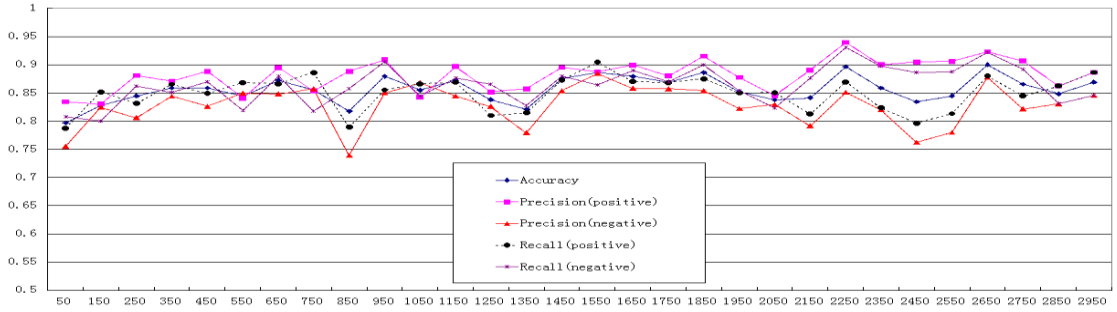

FIGURE 3 NB with presence representation under different feature dimensions

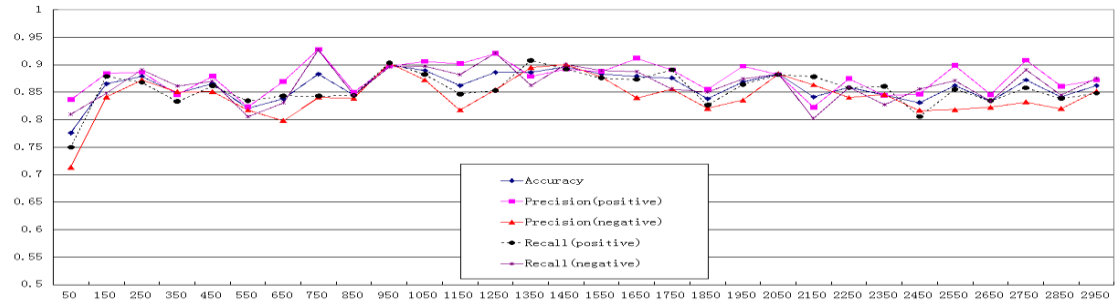

FIGURE 4 NB with TF representation under different feature dimensions

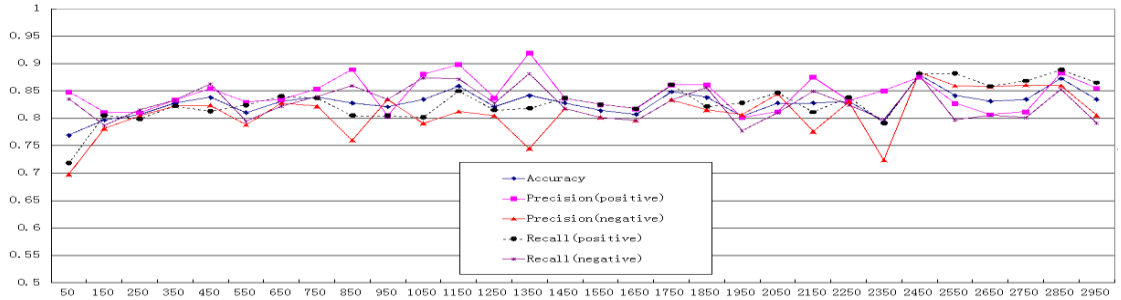

FIGURE 5 NB with TF-IDF representation under different feature dimensions

Figure 3 indicates the result of NB by accounting for feature presence, the most results of average accuracy are between 85% and 90%, the average accuracy peaks at 90% with 2650 features and the lowest average accuracy is 79.66% with 50 features. Figure 4 displays the result of NB with TF feature weight calculation method, the average accuracies distribute mainly between 85% and 90%, the top of average accuracy is 90% with 950 features, and the minimum average accuracy is 77.59% with 50 features. Figure 5 shows the results of NB with TF-IDF feature representation, the most average accuracies are between 80% and 85%, the peak of average accuracy is 87.93% with 2450 features and the minimum average accuracy is 76.90% with 50 features.

5.2 EXPERIMENT RESULTS OF ME CLASSIFIER

The performances of sentiment classification of ME with various feature representations and different feature dimensions are showed in Figures 6-8. Figure 6 indicates the result of ME by accounting for feature presence, the average accuracy peaks at 91.03% with 1250 (2350) features and the lowest average accuracy is 77.24% with 50 features. Figure 7 displays the result of ME with TF feature weight calculation method, the top of average accuracy is 91.38% with 2250 features, the minimum average accuracy is 74.83% with 50 features. Figure 8 shows the results of ME with TF-IDF feature representation, the peak of average accuracy is 91.72% with 1750 features and the minimum average accuracy is 80.00% with 50 features. As Figure 6-8 show, 90% results of the average accuracy of ME with three feature representations are above 85%.

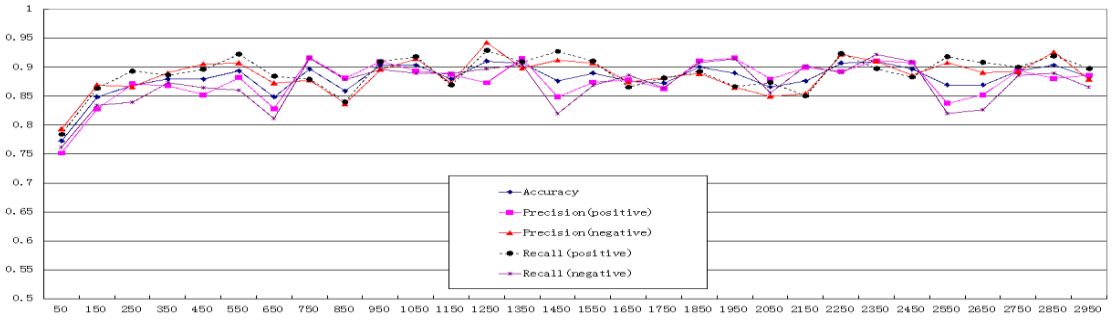

FIGURE 6 ME with presence representation under different feature dimensions

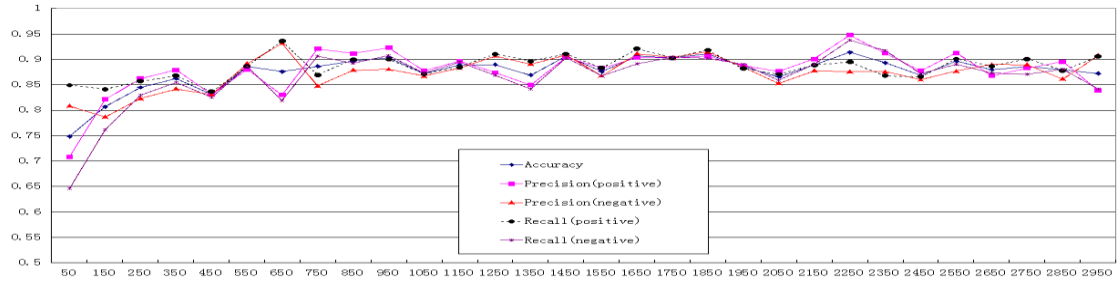

FIGURE 7 ME with TF representation under different feature dimensions

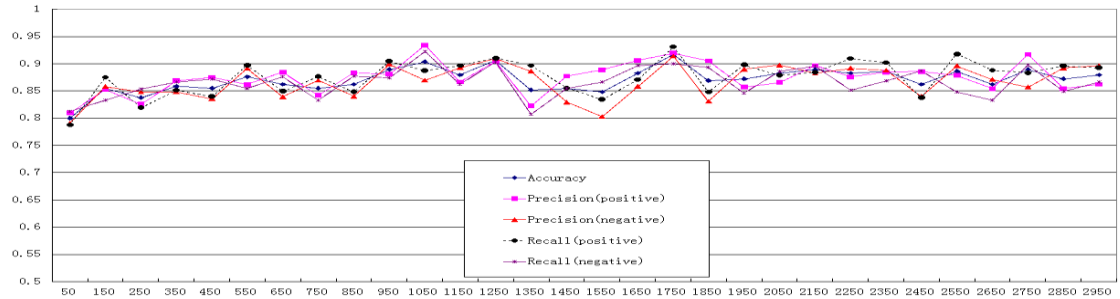

FIGURE 8 ME with TF-IDF representation under different feature dimensions

5.3 EXPERIMENT RESULTS OF SVM CLASSIFIER

The performances of sentiment classification of SVM with various feature representations and different feature sizes are showed in Figures 9-11. Figure 9 indicates the result of SVM by accounting for feature presence, the range of average accuracy is wide, the average accuracy peaks at 89.35% with 1850 features and the lowest average accuracy is 56.36% with 2250 features. Figure 10 displays

the result of SVM with TF feature weight calculation method, more than 90% results of the average accuracy are above 85%, the top of average accuracy is 87.97% with 350 features, the minimum average accuracy is 80.76% with 50 features. Figure 11 shows the results of SVM with TF-IDF feature representation, the most average accuracies are around 85%, the peak of average accuracy is 87.63% with 2550 features and the minimum average accuracy is 80.07% with 50 features.

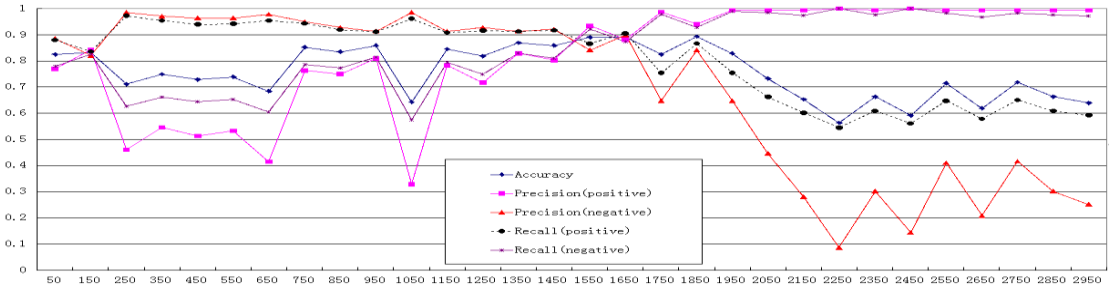

FIGURE 9 SVM with presence representation under different feature dimensions

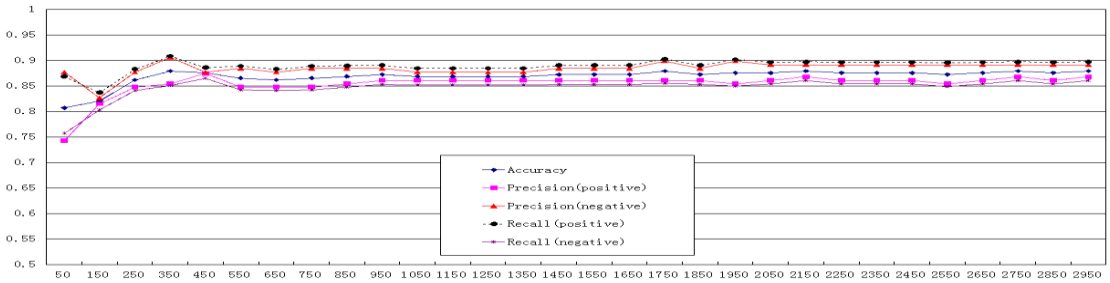

FIGURE 10 SVM with TF representation under different feature dimensions

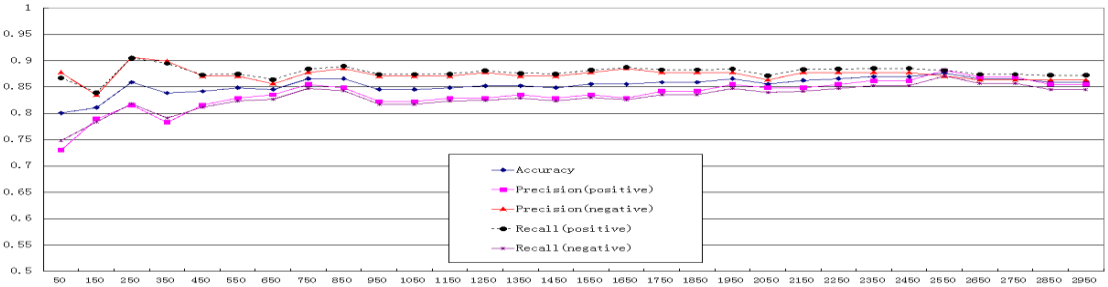

FIGURE 11 SVM with TF-IDF representation under different feature dimensions

5.4 COMPARISON AND ANALYSIS OF EXPERIMENT RESULTS

5.4.1 NB vs ME vs SVM

As Figures 3-11 show, the top of average accuracy will be varied according to different classifiers with different calculation methods of feature weight. Table 2 shows the statistical result of highest average accuracy for every classifier with three feature representations, ME is the best classifier for sentiment classification of online comments to news of commodity price fluctuation. The tops of average accuracy of ME with three feature representations are all above 91.00%, thereinto, the best average accuracy is 91.72% accounting for TF-IDF. ME slightly outperform NB. The highest average accuracy is 90% of NB with presence and TF feature weight calculation methods, but the number of features for NB with presence and TF feature representations are 950 and 2650, respectively.

TABLE 2 The top of average accuracy under three classifiers with three

	Presence	**TF**	**TF-IDF**
NB	90.00%	90.00%	87.93%
ME	91.03%	91.38%	91.72%
SVM	89.35%	87.97%	87.63%

Table 2 indicates that the minimum value of the top of average accuracies of three classifiers with three feature representations is 87.63%, which is achieved by SVM with TF-TDF. Table 3 shows the result of the number of features, if and only if the features are needed for three classifiers with different feature representations when their average accuracies are 87.63%. We find that ME achieve accuracy around 87.63% using few features than NB and SVM, especially ME with presence and TF-IDF feature weight calculation methods. NB requires fewer features than SVM with three feature representations.

TABLE 3 The feature dimensions when average accuracy achieve 87.63% in first time

	Presence	TF	TF-IDF
NB	950	250	2450
ME	350	550	950
SVM	1550	350	2550

The negative precision and positive precision of three classifiers with different feature weight calculation methods is varied. Table 4 displays the sum of absolute of D-value from negative precision and positive precision of three classifiers. For NB, with three feature representations, the positive precision is higher than negative precision. SVM achieves better negative precision. However, the sum of SVM with presence is more than others. Figure 9 shows the difference between negative precision and positive precision of SVM with presence is larger. The gap between positive precision and negative precision of ME is smallest.

TABLE 4 The sum of absolute of D-value from negative precision and positive precision

	Presence	TF	TF-IDF
NB	1.676	1.166	1.466
ME	0.932	0.951	1.002
SVM	12.197	0.909	1.165

The stability of three classifiers with three feature representations is different. Table 5 shows the mean values and variances of average accuracy. The variances of SVM with TF and TF-IDF is less than other variances, but the variance of SVM with presence is largest. Although ME outperform SVM and NB, its variance is larger. The variances of NB are smaller than the variances of ME. Thus, the result demonstrates a descending order of the stability of three classifiers as SVM (TF, TF-IDF)>NB>ME. The performance of SVM with presence is most unstable. In practice application, if you pay attention to the stability of classifiers, you can select SVM and TF to achieve higher average accuracy to sentiment classification of online news comments of commodity price fluctuation.

TABLE 5 The mean values and variances of average accuracy

	mean values of average accuracy (%)			variances of average accuracy		
	Presence	TF	TF-IDF	Presence	TF	TF-IDF
NB	85.62	86.06	82.77	0.0240	0.0265	0.0224
ME	88.0	87.66	87.10	0.0270	0.0336	0.0228
SVM	77.58	86.93	85.38	0.0993	0.0158	0.0160

5.4.2 Presence vs TF vs TF-IDF

For presence, TF and TF-IDF, combining with three classifiers to sentiment classification of online comments to news of commodity price fluctuation, the best average accuracy is achieved by TF-IDF and ME. However, different classifiers are suitable for different feature representations. NB with TF can use small size of features to achieve higher average accuracy. SVM with presence achieves the best average accuracy.

Table 3 shows that, three classifiers with different feature representations achieve the same average accuracy, they use a fewer features when they adopt TF feature weight calculation method than presence and TF-IDF. As Table 2 and Table 5 show, compared with other feature representations, TF has the superiority on the aspect of top of average accuracy and mean value of average accuracy. Table 4 displays that, the gap between negative precision and positive precision of TF is smallest. In this paper, TF is the best feature weight calculation method and its compatibility is best.

5.4.3 The number of features

As is shown above, the results of sentiment classification of online comments to news of commodity price fluctuation, except the accuracy of SVM with presence, which can use 50 features to achieve 82.47% accuracy, the average accuracy of three classifiers with different feature representations with low feature dimension is low. We find that the accuracies are low when classifiers with few features, but with the number of features increasing, the accuracies of three classifiers reach their peaks and then decline or fluctuate. For instance, the average accuracy of SVM with presence peaks 89.35% with 1850 features and then declines, the average accuracy of ME with presence peaks 91.03% with 1250 features and then fluctuates. This proved that, the effects of classifiers are influenced by feature dimensions when few features are used for classification, because the helpful features are not included. Moreover, after the accuracies reach their peaks, the effect of feature dimensions for classifiers is small, and the average accuracy can be improved through perfecting classifiers, feature weight calculation method and feature extraction method.

As Figures 2-11 show, according with the formula of recall and precision, the trend of positive recall and negative precision is same, the trend of negative recall and positive precision is same. What's more, the experiment result of recall and precision also proved the veracity of classifiers.

5.4.4 Model analysis

Overall, the maximum entropy classifier has outstanding performance for commodity price fluctuation news comments classification compared to naive Bayes and SVM. The deep reasons are related to the maximum entropy theory and character of these comments. The commodity price fluctuation news comments classification has the peculiarity that a lot of comments express the subjective opinion through the objective words, which enhance the difficulty of classification. However, the maximum entropy theory is based on that it does not make any hypothesis and remain maximum entropy for the unknown information. Hence, maximum

entropy classifier has strong robustness for our text data. Compared to two other classifiers, the theory of naive Bayes classifier is based on assumption that the features are independent with each other. In fact, the dependent relationship is existing between features. Hence, the classification result of naive Bayes is little worse than maximum entropy. For SVM classifier, the presence feature representation method has instability. When the feature dimensions is large enough, the classification accuracy reduce rapidly. However, the TF and TF-IDF feature representation methods for SVM have strong stability. The reason is that SVM classifier with presence feature representation will result in data sparse problem when the feature dimensions is very high, which make the accuracy reduce.

6 Conclusions

Since 2000, sentiment analysis has become a very active research area in linguistics and natural language processing. Although there are many researches about sentiment classification, most research is about product and service and little research had been done about sentiment classification of news comments. In this paper, we focus on the sentiment classification of online comments to news of commodity price fluctuation. We analyze the characteristics of Naive Bayes model, maximum entropy model and SVM model.

From the performance, we find that machine learning techniques perform quite well in the domain of sentiment classification of online comments to news of commodity price fluctuation. Comparing NB and SVM, ME is the most effective and efficient. The top accuracy of three classifiers is 92.72%, which achieves by ME with TF-IDF. However, SVM is the most stable classifier. With different feature representations, the accuracies of three classifiers reach their peaks. Considering feature dimensions simultaneously, TF is the best feature weight calculation method. The results of experiment proved that the effects of classifiers are affected if the number of features is small. With the increasing of dimensions of features, the influence of features dimensions is reducing. In the practical application, the dimensions of features should be chosen properly. Only in this way, can the result of classification will be efficient and accurate.

The value of this research is big, it can be useful for different areas, such as governments, business and so on. On the basis of this research, future research will be extended to the sentiment analysis on cloud platforms to solve all kinds of big data problems.

Acknowledgments

This work was mainly supported by National Natural Science Foundation of China (71363038), Natural Science Foundation of Inner Mongolia, China (2012MS1008, 2013MS1009), Scientific Research Project of Colleges and universities in Inner Mongolia, China (NJSZ12047).

References

[1] Pang B, Lee L, Vaithyanathan S 2002 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* **10** 79-86
[2] Turney P D, Littman M L 2003 Measuring praise and criticism: Inference of semantic orientation from association *ACM Transactions on Information Systems (TOIS)* **21**(4) 315-46
[3] Mullen T, Collier N 2004 Sentiment Analysis using Support Vector Machines with Diverse Information Sources *EMNLP 2004* **4** 412-8
[4] Kennedy A, Inkpen D 20006 Sentiment classification of movie reviews using contextual valence shifters *Computational Intelligence* **22**(2) 110-25
[5] Tang H, Tan S, Cheng X Q 2007 Research on sentiment classification of Chinese reviews based on supervised machine learning techniques *Journal of Chinese information processing* **21**(6) 88-94
[6] Devitt A, Ahmad K 2007 Sentiment polarity identification in financial news: A cohesion-based approach *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* 2007 984-91
[7] Tian F, Gao P, Li L, Zhang W, Liang H Qian Y Zhao R 2014 Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems *Knowledge-Based Systems* **55** 148-64
[8] Li W, Xu H 2014 Text-based emotion classification using emotion cause extraction *Expert Systems with Applications* **41**(4) 1742-9
[9] Guan W, Gao H, Yang M, Li Y, Ma H, Qian W, Cao Z, Yang X 2014 Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events *Physica A: Statistical Mechanics and its Applications* **395** 340-351
[10] Pang B, Lee L 2008 Opinion mining and sentiment analysis *Foundations and trends in information retrieval* **2**(1-2) 1-135
[11] Haney C 2014 Sentiment Analysis: Providing Categorical Insight into Unstructured Textual Data *Social Media, Sociality, and Survey Research* 35-59
[12] Balahur A, Mihalcea R, Montoyo A 2014 Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications *Computer Speech & Language* **28**(1) 1-6
[13] Balahur A, Turchi M Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis *Computer Speech & Language* **28**(1) 56-75
[14] Turney P D 2002 Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews *Proceedings of the 40th Annual Meeting on Association For Computational Linguistics* 417-24
[15] Hu M, Liu B 2004 Mining and summarizing customer reviews *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining ACM 2004* 168-77
[16] Ding X, Liu B, Yu P S 2008 A holistic lexicon-based approach to opinion mining *Proceedings of the 2008 International Conference on Web Search and Data Mining ACM 2008* 231-40
[17] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M 2011 Lexicon-based methods for sentiment analysis *Computational linguistics* **37**(2) 267-307
[18] Ye Q, Zhang Z, Law R 2009 Sentiment classification of online reviews to travel destinations by supervised machine learning approaches *Expert Systems with Applications* **36**(3) 6527-35
[19] Yu B 2008 An evaluation of text classification methods for literary study *Literary and Linguistic Computing* **23**(3) 327-43

[20] Zhang Z, Ye Q, Zhang Z, Li Y 2001 Sentiment classification of Internet restaurant reviews written in Cantonese *Expert Systems with Applications* **38**(6) 7674-82
[21] McCallum A, Nigam K A 1998 comparison of event models for naive bayes text classification *AAAI-98 workshop on learning for text categorization* **752** 41-8
[22] Lewis D D 1998 Naive (Bayes) at forty: The independence assumption in information retrieval *Machine learning: ECML-98* Springer Berlin Heidelberg 4-15
[23] Rish I 2001 An empirical study of the naive Bayes classifier *IJCAI 2001 workshop on empirical methods in artificial intelligence* **3**(22) 41-6
[24] Berger A L, Pietra V J D, Pietra S A D 1996 A maximum entropy approach to natural language processing *Computational linguistics* **22**(1) 39-71
[25] Chen S F, Rosenfeld R 2000 A survey of smoothing techniques for ME models *IEEE Transactions on Speech and Audio Processing* **8**(1) 37-50
[26] Nigam K, Lafferty J, McCallum A 1999 Using maximum entropy for text classification *IJCAI-99 workshop on machine learning for information filtering* **1**: 61-7
[27] Joachims T 1998 Text categorization with support vector machines: Learning with many relevant features *Springer Berlin Heidelberg*
[28] McCallum, Andrew Kachites 2002 MALLET: *A Machine Learning for Language Toolkit* http://mallet.cs.umass.edu
[29] Chih-Chung Chang and Chih-Jen Lin 2011 LIBSVM : *a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Authors

Yan Zhao, born in July, 1982, Tong Liao, Inner Mongolia, China

Current position, grades: Associate Professor at the College of Management, Inner Mongolia University of Technology.
University studies: Doctor's degree in Economics from Beijing Central University of Finance and Economics in June 2011.
Scientific interests: internet public opinion, information dissemination, management science and engineering, emotion recognition.
Publications: more than 12 papers.

Suyu Dong, born in October, 1989 ,Liang Cheng, Inner Mongolia, China

Current position, grades: student at the College of Management, Inner Mongolia University of Technology.
University studies: Bachelor degree in Software Engineering from East China Jiaotong University in June 2012.
Scientific interests: Management science and engineering sentiment analysis.

Jing Yang, born in May, 1981, Feng Zhen, Inner Mongolia, China

Current position, grades: Associate Professor at the College of Management, Inner Mongolia University of Technology.
University studies: Doctor's degree in Economics from Beijing Central University of Finance and Economics in June 2012.
Scientific interests: internet public opinion, information dissemination, emotion recognition.
Publications: more than 8 papers.