Sentiment Analysis for Microblog Related to Finance Based on Rules and Classification

Danfeng Yan[1st], Bo Hu[1st]
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing, China
yandf@bupt.edu.cn, jadfi@bupt.edu.cn

Jiafeng Qin[2nd]
State Grid Shan Dong Electric Power Research Institute
Jinan, China
qinjiafeng0118@163.com

***Abstract*—This paper proposes an approach based on rules and classification to analyse sentiment of Chinese microblogs related to finance. Firstly, we utilize an Improved Label Propagation Algorithm(I-LPA) to construct the sentiment lexicon automatically. Then according to microblogs' content, we divide them into multi-topic microblogs and single-topic microblogs by topic classification. As for multi-topic microblogs, rule-based sentiment analysis is applied. A three layers of filtering rule is used to identify the emotion agents of specified topic. Then we calculate the sentiment depending on syntactic dependency relationship between sentiment words and emotion agents. In other hands, single-topic microblogs exploit classification based on SVM to compute emotion. The experiments prove the validity of our methods. The results show that I-LPA is effective and the method of sentiment analysis is promising and outstanding for not only single-topic microblogs but also the multi-topic.**

Keywords—Rule-based sentiment analysis; I-LPA; SVM; emotion agents;

I. INTRODUCTION

As a new kind of interactive multimedia blog, Chinese microblog has been an indispensable social tool in people's lives. You can share your perspectives at any time by PC or smartphone. As a result, a large amount of texts are produced in the microblog platform every day. As the largest microblog platform in China, there are more than 130 million daily active users in Sina Microblog [1].

Chinese microblog has the following characteristics:(1)Microblog is mainly used to publish the user's point of view, making it very suitable for analyzing sentiment. Furthermore, the large number of microblog users makes the opinion more universal and referential. (2) Microblog's text is short. Chinese microblog specifies length of text must be less than 140 words, which makes it more difficult to extract features. Anyway, the microblog reveals great application value in the field of sentiment analysis even if it is challenging.

Though most of information in Chinese microblog talks about current hot spots and entertainment news, there are still many microblogs related to finance mixed in massive texts. Considering that this field is too professional, the information is always ignored by researchers. In this paper, we aim at sentiment analysis for microblogs related to finance, which can help investors and government officials make decisions.

The concept of sentiment computing was first proposed by Professor R.Picard [2] in 1997. It was defined as "Computing that relates to, arise from, or deliberately influences emotions". Affective computing has always been a hot research topic in artificial intelligence. As the main carrier of network information, more and more researchers pay attention to emotional analysis in texts. The main task of text sentiment analysis is to classify the textual data into three categories: positive, negative and neutral. Positive and negative emotional tendencies attract most researchers' attention, so sentiment analysis can also be seen as two-polarity problem. This paper first only pays attention to positive emotion and negative emotion and then extends the proposed method to analyze sentiment for three-polarity.

In the past work, most researches are devoted to computing the emotion of the whole microblog because there are usually only one topic in the microblog. However, microblog users may discuss several kinds of financial sectors, which form multiple topics in one piece of microblog. So we propose a method to take different measures to deal with the single-topic microblogs and multi-topic microblogs. We use traditional supervised method to measure the sentiment of single-topic microblogs and propose new unsupervised rules to analyze the multi-topic microblogs.

The main contributions of this paper are: 1) analyze the sentiment of multi-topic microblogs and single-topic microblogs using different algorithms. A rule-based method is proposed to deal with the multi-topic microblogs and classification based on SVM is employed for single-topic microblogs; 2) improve the traditional Label Propagation Algorithm(LPA) to set up sentiment lexicon in finance; 3) demonstrate the necessity of building sentiment lexicon in sentiment analysis. And this paper establishes different lexicons for sentiment analysis of multi-topic and single-topic microblogs.

The rest of this paper is organized as follows: Section II will describe the related work of sentiment analysis. Section III will elaborate on the principles of methods and the design of algorithms in computing emotion. In section IV, we conduct experiments to illustrate the validity and scalability of our method. The results and related analysis will be shown detailedly. The last section concludes our work and the prospects for the further work are presented.

2375-9356/18/$31.00 ©2018 IEEE
DOI 10.1109/BigComp.2018.00026

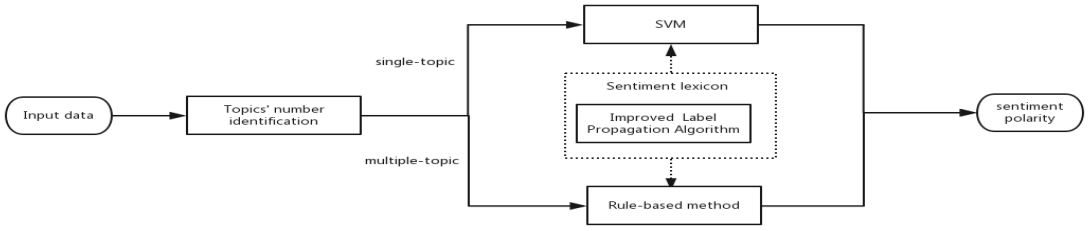

Figure 1. Flow chart of sentiment analysis

II. RELATED WORK

Text sentiment analysis involves several technology including natural language processing, features extraction, text classification and so on. At present, there are two main ways in emotion analysis. One is rule-based method, which takes advantage of syntactical structure, dependency relation and rules of grammars. The other one is based on machine learning.

[3] describes a rule-based sentiment analysis algorithm for polarity classification of financial news articles. Sentiment composition rules are introduced to tackle different kinds of phrase sentiment compositions. Take composition of verb-noun and noun-verb for example, the combination of a positive noun phrase with a negative verb phrase results in a negative noun phrase. A total of six combinations are discussed by the author. [4] designs an approach based on Apriori algorithm to extract attributes related to finance. Then calculating the emotional intensity of emotion analysis unit from financial texts that contain extracted attributes according to the predefined semantic rules. These rules focus on the position of negative modifiers. However, both of them are only suitable for the analysis of a single topic.

Additionally, sentiment analysis can be seen as a classification problem. Silva [5] employs classifier ensembles to compute the emotion for tweets. The classifier is formed by diversified components, including Naive Bayes, SVM, Random Forest and Logistic Regression. Also, different information sources are discussed, such as textual data, emoticons and lexicons. Lu [6] determines the sentiment polarity of short Chinese texts based on SVMs. the weight of sentiment words and words next to the sentiment words are enhanced by improving tf-idf value. [7] proposes a variant of the standard CHI algorithm to make emotion words more easily to be selected as the feature words. Zhao [8] combines rules and classification with linear weighting, obtaining their respective advantages. Besides classification, topic model can also be adopted in sentiment analysis. [9] adds an emotion layer to LDA and adds micro-bloggers' social relation to achieve synchronized detection of sentiment.

As the fundamental of affective compute, construction of sentiment lexicon is necessary. There are some authoritative general sentiment dictionaries widely used by researchers, like NTUSD and Sentiment Lexicon Ontology [10]. Work from [11] extends the basic dictionary on the basis of the similarity of words. Turney proposes classical SO-PMI [12] algorithm, which determines the polarity by calculating the closeness between the word to be detected and the seed words. Graph-based algorithm can also be available for dictionary constructing. PageRank algorithm is used by [13] to iteratively compute the sentiment value of words.

III. MICROBLOG SENTIMENT ANALYSIS

In this paper, there are three steps to analyze sentiment of microblogs as shown in Fig.1. 1) Building a sentiment lexicon in finance by I-LPA. 2) Identifying the topics' number of microblogs. Topic classification is exploited for the identification [14]. So topics are identified automatically without manual work. Then microblogs will be divided into the multi-topic and the single-topic according to the number of topics. 3) Dealing with the single-topic and multi-topic microblogs using different algorithms. Rule-based method is utilized to calculate sentiment polarity of each topic in the multi-topic microblog. As for the single-topic microblog, it will be classified into two kinds of emotion polarities by SVM, namely positive emotion and negative emotion.

The reason for using different algorithms is related to the number of features. Fewer features in the multi-topic microblog lead to bad performance in general machine learning like SVM because of under-fitting phenomenon. Thus we present new method based on rules and it fits better. As for single-topic microblogs, all features in the text can be considered to describe the single topic, which is more suitable for classification. We will discuss it in detail in next section.

A. Sentiment lexicon construction based on I-LPA

SO-PMI [12] algorithm is a classical method in sentiment lexicon construction. In traditional SO-PMI algorithm, the degree of similarity between words is calculated by computing the co-occurrence of words in the large corpus. The disadvantage is that SO-PMI only considers the relation between the seed words and unknown words. But unknown words can also learn the emotion information from each other. So we adopt Label Propagation Algorithm to construct sentiment lexicon and improve it.

Original LPA is proposed by Raghavan [15] to detect community structures. The algorithm is popular for its simple implementation and fast convergence. However, LPA updates the label in a random way when there are more than one candidate labels. LPA will choose the label randomly when the number of neighbour with positive label and negative label are the same. Also, the order of nodes to be updated is uncertain. Therefore, we overcome these problems and combine with

PMI similarity to get a better performance in our work. The basic steps of our improved algorithm are as follows:

Step1: Select 10 pairs of positive and negative words manually as the seed words like SO-PMI algorithm does. The seed words should be typical sentiment words in finance. And we denote the set of seed words as S.

Step2: Establish the graph G=(V,E,W,L,C) based on the word relations. Vertex set V is made up of words in the corpus, including seed words and non-seed words. All non-seed words are unlabelled. Besides, only verbs and adjectives are considered. If v_i and v_j appear in the same sentence, there is an edge e_{ij} between the two nodes. Edge weight w_{ij} is represented by normalized PMI value between v_i and v_j. Label set L={ $l_1, l_2, .., l_i$ },where l_i represents the sentiment polarity of v_i. Besides, each word v_i has a confidence set to indicate the probability of which sentiment polarity it belongs to. It is denoted as c_i and $c_i=(p_i,n_i)$, where p_i means confidence of positive emotion and n_i represents negative confidence. At the initial stage, the label of all non-seed words are initialized to 0. And the confidence set are assigned to (0,0).

Step3: Sort the non-seed words. Scan all non-seed words and calculate the relevance of each word to seed words. Relevance R_i of non-seed word v_i is defined as (1)

$$R_i = \sum_{j \in S} w_{ij} \tag{1}$$

R_i indicates the impact of the seed words on v_i. So we sort the non-seed words by the relevance from high to low.

Step4: Start to propagate the label of seed words to non-seed words in graph G with the sorted order. For each non-seed word v_i, we update c_i as (2):

$$c_i = (\sum_{j \in N} p_j * w_{ij}, \sum_{j \in N} n_j * w_{ij}) \tag{2}$$

N is the set of neighbor nodes of v_i . Then we calculate the label l_i based on confidence by (3):

$$l_i = \begin{cases} 1 & p_i - n_i > \theta \\ 0 & |\, p_i - n_i | < \theta \\ -1 & p_i - n_i < -\theta \end{cases} \tag{3}$$

θ is the preset confidence threshold. Note that only when l_i changes does the c_i update.

Step5: Repeat step 4 until the labels of all non-seed words are stable.

In this paper, we exploit I-LPA to construct financial sentiment lexicon. We improve stability of label updating by introducing confidence and avoid the random update order by sorting the non-seed words.

B. Rule-based Method for multi-topic microblogs

In sentiment analysis of multi-topic microblogs, we denote the sentiment polarity of ith topic as T_i. And the sentiment polarity of the multi-topic microblog is denoted as a tuple $M=[T_1,T_2,\ldots T_n]$, where n is the number of topics. So we focus on how to calculate the value of T_i. There are three steps in calculating sentiment polarity in one topic. 1) Extracting relevant emotion agents. 2) Parsing dependency relation of the sentence and set rules to compute the sentiment value of each agent on basis of dependency. 3) Accumulating sentiment value of each agent to get the final sentiment polarity of the topic. We will describe the details in this section.

1) Extraction of emotion agents: Emotion agents represent the holders which express the sentiment. In this paper, we limit ourselves to those agents related to financial attributes. We first extract all nouns as candidate agents and then acquire the financial emotion agents by designing a three-layer filtering rule. The order of layers arrangement is based on the confidence from high to low. Prior to filtering, seed agents are selected to indicate which financial attributes we care about. For example, if a financial topic is about metal concepts, the seed agents can be set as "metal" and "price of metal" because we only concentrate on those emotion agents which are associated with the metal and its price. Then our filtering rules will discard the irrelevant agents. At last, the agents describe the metal or the price of metal will remain, like "gold", "silver", "price of gold", "price of silver" and so on. The flow chart is shown as Fig.2.

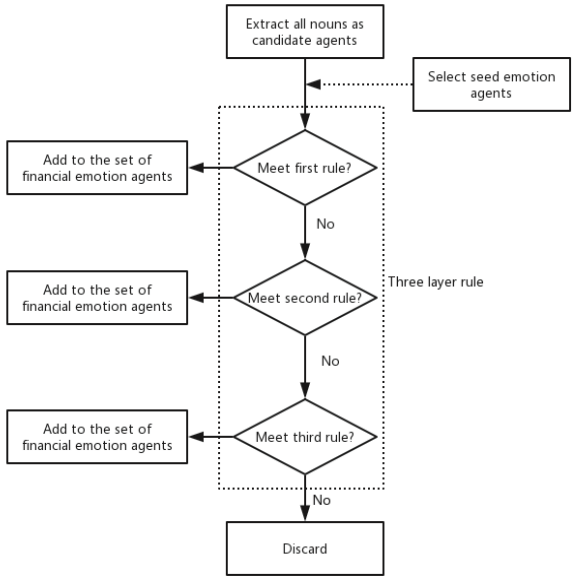

Figure 2. Flow chart of extraction of financial emotion agents

Our first filtering rule is based on inclusion relation of phrases. It is considered to find out the noun combination forms. If the phrase contains a seed agent, it meets the rule. For instance, "People's Bank" will be selected if "Bank" is a seed agent in the topic about banking.

Second rule is based on Hownet [16] which is a knowledge base describing the relationship among Chinese words. Hownet uses a tree to present the semantic relation and the similarity can be measured by the path distance between words. We adopt the method proposed by Liu [17] to calculate the similarity among the candidate agents and seed agents based on

Hownet. We only remain the agents with similarity exceeding the preset threshold.

The last rule utilizes the tool Word2Vec [18]. It can map the word into vector by training a large corpus. The similarity between words can be calculated by cosine similarity between vectors. We also remain candidate agents by similarity filtering.

Note that emotion agents also have modified polarity. When a positive sentiment word qualifies a positive agent, the phrase expresses positive emotion. Otherwise it is negative. So we also specify the modified polarity in selecting seed agents and the modified polarity of remained emotion agents are consistent with the corresponding seeds. Although the extraction of agents needs a little seeds, it can be applied to any field as a semi-supervised method.

2) Rules of affective computing for emotion agents: Taking advantage of above rules, emotion agents with modified polarity will be selected. Then we set rules to compute the sentiment value of each agent. The core idea is to accumulate the sentiment value of all sentiment words which modify the emotion agent.

Firstly, we use Stanford Parser [19] to parse the microblog. Stanford Parser is tool to show the dependency relation between words. As a result, we can establish the dependency graph D=(V,E). V represents the set of words appearing in the microblog. Edge e_{mn} will be set up if there exists the modified dependency between word v_m and word v_n. Also, we denote the path from v_m to v_n as P(m,n).

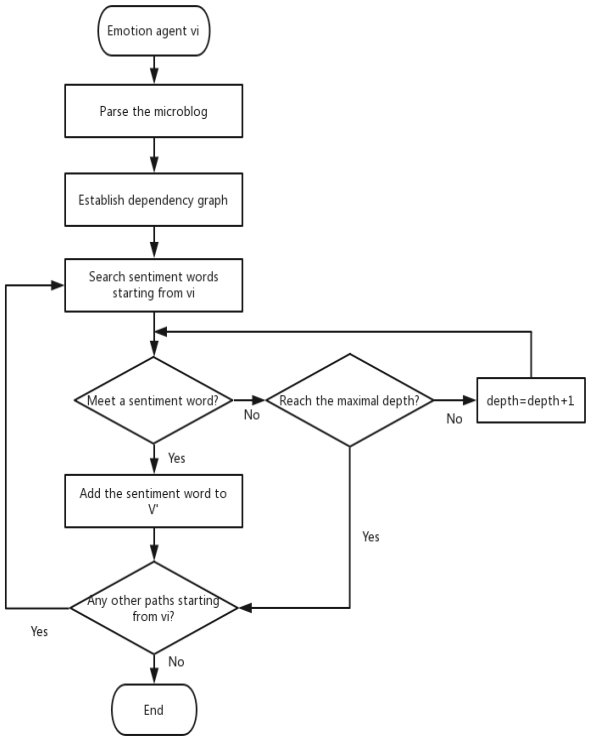

Figure 3. Work flow of computing rules for emotion agents

For an emotion agent v_i, we denote its sentiment polarity as A_i. Our target is to find out all sentiment words $V'=\{v_1,v_2,..v_j\}$, which are used to modify the emotion agent v_i. If a sentiment word v_j acts on the agent v_i, there must exists P(i,j). We search for the sentiment words by depth-first searching. There are two rules in search process. The first rule is that the searching depth should be limited. We set the parameter to 4, which indicates that the maximal length of a path is no bigger than 4. Also, we set the second rule that the depth-first searching returns when it meets the first sentiment word in one path, which means there is no other sentiment words in P(i,j) except for v_j. According to above two rules, we can find all P(i,j) that satisfies the requirements and get all sentiment words $v_j \in V'$. The work flow is illustrated by Fig.3.

For each sentiment word $v_j \in V'$, we employ the same rules to find out the negative modifiers of it. We denote the sentiment polarity as Sv_j and the number of negative modifiers of v_j is denoted as K. So Sv_j is calculated as (4), where l_j is the sentiment label calculated in I-LPA.

$$S_{vj} = \begin{cases} l_j & K\%2 = 0 \\ -l_j & K\%2 = 1 \end{cases} \tag{4}$$

The maximal depth for searching privative words is set to 1. It is adjusted through a considerable amount of experiments. The searching depth for sentiment words seems larger because all words in one path have remained including meaningless prepositions and conjunctions. As for negative modifiers, they usually follow the word modified closely.

In the end, the value of A_i is calculated as bellows:

$$A_i = \sum_{vj \in V'} S_{vj} \tag{5}$$

3) Rules of affective computing for the topic:

We get the sentiment value of each emotion agent in one topic from previous step. A_{ij} means the sentiment value of jth emotion agent in ith topic. We accumulate the value of emotion agents and normalize the result. The formula is shown as (6), where k is the number of emotion agents in ith topic.

$$T_i = \begin{cases} 1 & \sum_{j=1}^{k} A_{ij} \geq 0 \\ -1 & \sum_{j=1}^{k} A_{ij} < 0 \end{cases} \tag{6}$$

C. *Sentiment classification for single-topic microblogs*

As for single-topic microblogs, we exploit traditional classification to classify the sentiment polarity. We select Support Vector Machine(SVM) as the method to determine the sentiment polarity. We choose SVM because it has a good generalization ability, making it one of the most widely used and the best classifiers. Linearly separable data can be well analysed with SVM. As for linearly inseparable problem, it will also be converted to linearly separable one. By using the non-linear mapping algorithm, the linear separable samples of

low dimensional input space are transformed into high dimensional feature space.

For the sake of simplicity, we only select sentiment words appearing in microblogs as features to explore the influence of sentiment words. Let $D=[f_1,f_2,..f_t]$ expresses the feature vector of the text, where f_t is the vector value of tth sentiment word Similarly, we use depth-first algorithm to search for the negative modifiers for each f_t . Then f_t is calculated like Sv_j as (4).

IV. EXPREIMENTS AND RESULTS ANAIYSIS

We design four experiments to analyze and describe the selection of different methods and the improvement of lexicon in sentiment analysis.

The first experiment proves that our proposed rules are more appropriate for multi-topic microblogs than general SVM and the conclusion is opposite in single-topic microblogs. We adopt sentiment lexicon constructed by I-LPA for the first experiment because the basic sentiment lexicon or their combinations leads to the same results. Experiment 2 is designed to compare the effects of different sentiment lexicons on the sentiment analysis. Then experiment 3 is conducted to analyse the influence of I-LPA on the construction of sentiment lexicon. The last experiment extends two-polarity sentiment to the three-polarity to further identify the neutral emotion.

Here grid search algorithm is brought into the selection of the best parameters of SVM. At last linear kernel is chosen. We use 10-fold cross-validation to calculate the micro F-scores. In all figures of results, the horizontal axis represents the confidence threshold. The higher the threshold, the smaller size of the sentiment lexicon. And the vertical axis represents the micro F-scores.

A. Experimental data

About 700 thousand microblogs are collected from Sina Microblog. We filter out the microblog that is irrelevant to finance to improve the accuracy of PMI computing.

TABLE I. DISTRIBUTION OF MULTIPLE TOPICS

Topic	Count
Metal concepts	225
Banking industry	174
Petroleum industry	222
Currency	183

TABLE II. DISTRIBUTION OF SINGLE TOPIC

Topic	Count
Metal concepts	310
Petroleum industry	284
Banking industry	212
Insurance industry	242

TABLE III. EXAMPLES OF EXPERIMENTAL DATA

Microblogs	Topic	Sentiment label
I think gold has room to go up and it can continue to be held. Also, I am not optimistic about the prospects of dollar.	Metal concepts; Currency	+1;-1
New China Life Insurance has a strong performance and can be held for a long time.	Insurance industry	+1

We have constructed finance dictionary in advance to help word segmentation. Therefore, if the number of financial words contained by the microblog is less than a threshold, it will be deleted.

Finally we acquire about 55 thousand microblogs. In view of the informal text format, we remove expression symbols and urls. With regard to microblogs' sentiment labelling, there are 523 microblogs with multiple topics and 1048 microblogs with single topic. The work is done by four people together. Table I and Table II show the topic distribution of multi-topic and single-topic microblogs. Table III gives the examples of sentiment labelling for multi-topic and single-topic microblogs.

In addition, a basic sentiment lexicon should be created to do the contrast experiment. We choose Sentiment Lexicon Ontology [10] labelled by Dalian University of Technology and there are 11196 positive words and 10755 negative words. Besides, we manually create a privative dictionary containing 43 words.

B. Experiment 1:method comparison between rules and SVM

In this experiment, we apply rule-based method and SVM to the sentiment analysis of multi-topic microblogs and single-topic microblogs. Then we discuss the effects.

1) Performance of rule-based method and SVM in multi-topic microblogs

Firstly, we concentrate on the multi-topic microblogs. The result is shown as Fig.4.

In the Fig.4, the range of confidence threshold is from 0.005 to 0.05 because the experiments get the best performance in the range. For rule-based method, we can see the result achieves the best balance when the threshold is 0.01.

Obviously, rules get better performance than SVM in multi-topic microblogs. The highest micro F-scores increases from 81.95% to 84.58%. The reason is that machine leaning occurs under-fitting phenomenon when the feature words are too few while rule-based algorithm doesn't. However, the result of SVM is not very bad though each specified topic only has one feature word on average. This is because few features indicate that each feature word gets more valid information and it is crucial. It also proves that the sentiment lexicon we constructed is good. Because high micro F-scores result from the high accuracy of identification of the sentiment words, which are expanded automatically.

Also, few features indicate that the specified topic is difficult to appear the affect-transfer problem. So simple rules gain effective results.

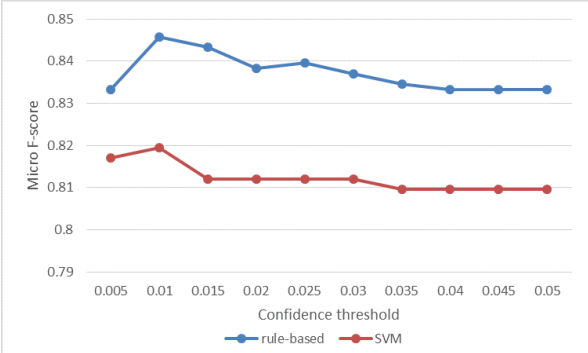

Figure 4. Method comparison in multi-topic microblogs

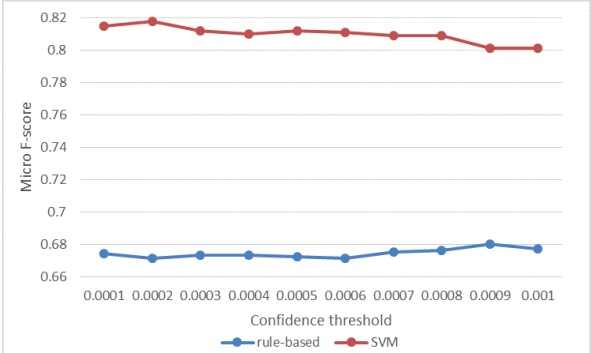

Figure 5. Method comparison in single-topic microblogs

Fig.4 shows that the change of micro F-scores is smooth. This is related to the characteristics of multi-topic microblogs. Multiple topics mean that the expression ability of each topic is limited. Hence, the authors usually use the most professional and precise words to express their sentiment about finance. Nonfinancial noise words rarely occur even if they are introduced, owing to the phenomenon of excessive concentration of words. Therefore, as the scale of lexicon expands, micro F-scores don't decrease rapidly.

Furthermore, SVM algorithm is supervised so it needs more time to label and train the datasets. The experiment indicates that the general classification method like SVM is not fit for the sentiment analysis of multi-topic microblogs. In contrast, our proposed rules are effective.

2) Performance of rule-based method and SVM in single-topic microblogs

We can see from Fig.5 that in single-topic microblogs classification method is better than rules. When the whole microblog discuss only one topic, the syntactical structure will be complex. So the emotion becomes hard to be identified with rules. Meanwhile, the increased features make it suitable to use method of machine learning. We can see that the highest accuracy is increased from 68% to 81.75%. And when the threshold exceeds 0.0002 the performance starts to degrade slowly with the introduction of noise features.

Notice that the best confidence threshold is smaller than that in multi-topic microblogs. It indicates that larger sentiment lexicon makes better impact in SVM algorithm. The reason is that more valid feature words lead to better performance, especially for short texts. In this case, construction of the sentiment lexicon can be regarded as a method of feature extraction. The words in sentiment lexicon is either sentiment words or items which are closely related to sentiment words. And a wider variety of words are used in expression in single-topic microblogs, so the probability that the words in lexicon are selected as features becomes higher.

Similarly, the curves varies smoothly when the scale of lexicon expands. The number of features increases much in single-topic microblogs. So the tolerance for few errors increases too. The errors are caused by the false identification of sentiment words when using I-LPA.

Hence we choose rules to analyze sentiment for multi-topic microblogs while we select SVM to classify the emotion for single-topic microblogs.

C. Experiment 2: comparison between financial sentiment lexicon and the basic lexicon

In finance, the sentiment words are tend to be unique and professional, so we need to compare our sentiment lexicon with basic Sentiment Lexicon Ontology to verify the validity. In this experiment, rule-based method is used for sentiment analysis in multi-topic microblogs and SVM is exploited to computing the emotion of single-topic microblogs, because the first experiment proves that it produce the best results.

The results from Fig.6 and Fig.7 show that our domain sentiment dictionary gets the best performance in either rule-based method or SVM algorithm. Note that when the polarity of a word meets conflict in financial sentiment lexicon and basic dictionary, we take financial sentiment lexicon as standard.

We can declare that the sentiment words in the basic lexicon is not fit for the financial sectors. Characteristic of financial texts is that there exists strong domain correlation. More noises are introduced when combining the basic sentiment lexicon and financial sentiment lexicon. In addition, comparing with the performance of combination lexicon in Fig.6 and Fig.7, we can see that sentiment analysis for the single-topic microblogs is less sensitive to noises than the multi-topic microblogs because the performance of the combination dictionary is closer to the performance of financial lexicon in Fig.6.

Fig.6 and Fig.7 suggest that the micro F-score is lowest when only using the basic sentiment lexicon. The F-score remains unchanged because the lexicon is static. The result seems not bad because the features are so few that SVM classifies all test sets as one class.

In conclusion, we only select the domain sentiment lexicon constructed by the I-LPA to analyze emotion of multi-topic microblogs and single-topic microblogs.

Finally 190 positive words and 196 negative words are integrated into the domain sentiment lexicon for multi-topic microblogs. The sentiment lexicon for single-topic microblogs

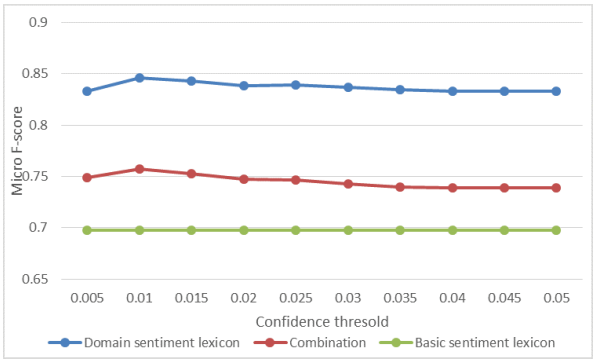

Figure 6. Lexicon selection in multi-topic microblogs

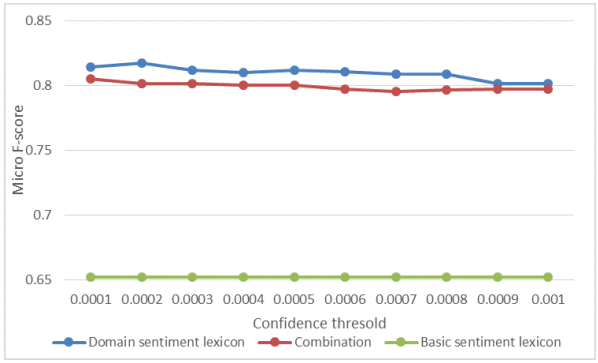

Figure 7. Lexicon selection in single-topic microblogs

contains 2215 positive words and 2023 negative words. Error tolerance is low for the method based on rules. If a positive financial sentiment word is identified as a negative word, rules are not able to compensate for the error. And the ultimate calculation is sensitive to the single word error in view of the little quantity of features. Also, multi-topic microblogs usually focus on the professional financial words. Because it contributes to express each topic more accurately with a few words. Therefore, the sentiment lexicon is small but precious enough in analyzing the multi-topic microblogs. As for classification, the number of useful features dominate the performance. The more import point is that classification is supervised. To some extent, training sets labelled manually can make up for mistakes in recognition of emotional words. So the sentiment lexicon is large.

D. Experiment 3: comparison between I-LPA and the traditional LPA

This paper designs an Improved Label Propagation Algorithm to construct the financial sentiment lexicon by the semi-supervised method. Then we do experiments to demonstrate that our algorithm outperforms original algorithms.

Note that the result of original algorithm is volatile on account of the randomness. Thus we calculate the mean of three experimental results as the final result.

Table IV indicates that I-LPA makes better performance than the traditional algorithm. One of the main reason is the greater local density of corpus data. We initialize the label to 0 for all unknown words. Thus a large number of unknown nodes will be clustered because of the high cohesion. When the quantity of labelled seed nodes is not large enough, chances are that the positive label or negative label is difficult to spread to the cluster, leading to lots of neutral labels in results.

TABLE IV COMPARISON IN METHODS OF LEXICON CONSTRUCTION

	Rule-based method in multi-topic microblogs	***Classification method in single-topic microblogs***
LPA	83.08%	65.23%
I-LPA	84.58%	81.75%

In the improved algorithm, we bring in the confidence to calculate the label for a node instead of exploiting the most frequently used label in the neighbours' nodes. Our method expands the influence of labelled nodes on unlabelled nodes by spreading the confidence. Even if the label of a node is neutral, it still carries the positive or negative confidence which can make a difference to its adjacent nodes.

Our method removes the randomness of result while improving the micro F-scores. The comparison shows that the micro F-score of rule-based method increases by about 1.5% while the micro F-score of classification increases by about 16.5%. This is because the classical algorithm only produces 30 negative sentiment words and 26 positive sentiment words owing to the poor capability of label propagation.

Experiment 4: expansion of neutral emotion

Particularly, we introduce neutral emotion to investigate the identification ability for three-polarity sentiment. We discuss the neutral emotion in two situations. If the microblog involves no sentiment words, it is neutral. Extra zero vectors are added to test the performance in this case. From Table V we can see that rules can identify all neutral labels. However, some positive and negative labels will be calculated as 0 falsely, lowering the F1-score of the all data. We can also learn it from Table V that reduplicate zero vectors can also be handled well by SVM.

We only adopt the recall score to measure the identification power for neutral sentiment. The main reason is that precision score is tend to be affected by the number of unbalanced samples. When positive labels or negative labels dominate the samples, the precision score of neutral label is sensitive to the labels which are computed falsely as neutral.

TABLE V RESULTS FOR THREE-POLARITY SENTIMENT ANALYSIS

	Rule-based method in multi-topic microblogs		***Classification method in single-topic microblogs***	
	Neu_recall	***All_F1***	***Neu_recall***	***All_F1***
Adding neutral emotion containing no sentiment words	100%	69.21%	100%	79.58%
Adding neutral emotion containing sentiment words	59.1%	63.02%	27.67%	67.46%

In second case we introduce neutral multi-topic and single-topic microblogs containing sentiment words. The quantity is equal to the first case. Obviously, the results from Table V indicate that both rule-based method and classification algorithm work worse compared with the first case.

It is clear that the classifier is confused when positive and negative features appear simultaneously. So it is hard to identify the neutral labels while it is easy to mislead the identification of positive and negative labels, resulting in the rapid drop in recall score. As for rule-based method, the recall score is almost 60% and the total F1-score achieves 63%, which is acceptable.

We can declare that rules adapt much better than SVM on the identification of neutral labels in the second case. If the number of positive sentiment agents and negative sentiment agents are the same in one microblog, it is labelled as 0 with higher probability in training sets from our observation, which is consistent with our rules. But SVM is not able to distinguish this linear relation.

In sum, the result suggests that rule-based method has good adaptability for sentiment analysis of three-polarity. There is still much room to improve for the classification method when the neutral microblogs contain sentiment words.

V. CONCLUSION

In summary, this paper presents a novel method to analyze the sentiment of microblog based on rules and classification. We construct the sentiment lexicon in the field of finance by improving traditional LPA. Then we deal with multi-topic microblogs and single-topic microblogs in different ways. We set rules to identify the polarity of sentiment for multi-topic microblogs because of the few features. A three-layer rule is adopted for the extraction of sentiment agents. For single-topic microblogs we utilize SVM to give out the result of classification.

We conduct four experiments and prove the validity of our method. There are four main conclusions from the results: 1) The rule-based method is simple, but is more effective and fit for the sentiment analysis of multi-topic microblogs than general classification that is used mostly at present. The disadvantage is the poor immunity for noises. As for the single-topic microblogs, the supervised classification on the basis of SVM performs best. It is stable owing to the better tolerance for noises, though the manually labelling is necessary. 2) Our financial sentiment lexicon substantially improve the performance compared with the basic dictionary. Also, we build up different lexicon for multi-topic and single-topic microblogs. 3) Traditional LPA has the disadvantages of strong randomness and weak propagation ability. Our proposed I-LPA can overcome the shortcomings by introducing the confidence and sorting the unlabelled words. 4) Our method can also be applied to analyze the sentiment for three-polarity. But there is still much room for improvement.

We believe that future research could focus on the improvement for three-classification.

ACKNOWLEDGMENT

This paper is supported by “National 863 project(No.2015AA050204)” and “National technology project(No. 520626170011)”. We would like to thank Sina for sharing microblog data, and Dalian University of Technology for providing the basic sentiment dictionary. We also thank the reviewers for their valuable comments.

REFERENCES

[1] Report on Development of Microblog Users.[Online]. Available: http://data.weibo.com/report/reportDetail?id=346

[2] R.W.Picard. Affective Computing[M]. Cambridge,MA;MIT Press,1997.

[3] L.I.Tan,W.S.Phang, K. O. Chin and P. Anthony, "Rule-Based Sentiment Analysis for Financial News," 2015 IEEE International Conference on Systems, Man, and Cybernetics, Kowloon, 2015, pp. 1601-1606.

[4] Wu Jiang,Tang Changjie,Li Taiyong,CUI Liang,"Sentiment analysis on Web financial text based on semantic rules,"Journal of Computer Applications,2014,02:481-485+495.

[5] Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka. "Tweet sentiment analysis with classifier ensembles." Decision Support Systems 66 (2014): 170-179.

[6] L.Xing, L.Yuan, W.Qinglin and L. Yu, "An approach to sentiment analysis of short Chinese texts based on SVMs," 2015 34th Chinese Control Conference (CCC), Hangzhou, 2015, pp. 9115-9120.

[7] D.Yuan, Y.Zhou, R.Li and P.Lu, "Sentiment analysis of microblog combining dictionary and rules," 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, 2014, pp. 785-789.

[8] B.Zhao,Y.He,C.Yuan and Y.Huang, "Stock market prediction exploiting microblog sentiment analysis," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 4482-4488.

[9] Huang FL, Yu G, Zhang JL, Li CX, Yuan CA, Lu JL. “Mining topic sentiment in micro-blogging based on microblogger social relation.” Ruan Jian Xue Bao/Journal of Software, 2017,28(3):694707 (in Chinese). http://www.jos.org.cn/1000-9825/5157.htm

[10] Sentiment Lexicon Ontology.[Online]. Available: http://ir.dlut.edu.cn/EmotionOntologyDownload

[11] Yang Chao,Feng Shi,Wang Da_ling,Yang Nan,Yu Ge,"Analysis on Web Public Opinion Orientation Based on Extending Sensiment Lexicon,"Joumal of Chinese Computer Systems.2010

[12] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classfication of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. 417−424

[13] Li Rong-Jun, Wang Xiao-Jie, Zhou Yan-Quan. “Semantic orientation computing using Page Rank model”, Journal of Beijing University of Posts and Telecommunications, 2010,33(5): 141−144

[14] Zheng Wei, Wang Chao-Kun, Liu Zhang, Wang Jian-Min. “A Multi-Label Classification Algorithm Based on Random Walk Model,” Chinese Journal of Computers. 2010,(08):1418-1426.

[15] Raghavan U N,Albcrt R,Kumara S. “Near linear time algorithm to detect community structures in large-scale networks”. Physical Review E,2007,76(3):036106

[16] Hownet. [Online]. Available: http://www.keenage.com/html/c_index.html

[17] Liu Qun,Li Sujian,"The calculation of lexical semantic similarity based on HowNet," The 3rd Chinese Lexical Semantic Workshop,2002:59-76

[18] Mikolov T, Chen K, Corrado G, et al. “Efficient estimation of word representations in vector space”. arXiv preprint arXiv:1301.3781, 2013:1-12.

[19] Stanford Parser. [Online]. Available:https://nlp.stanford.edu/software/lex-parser.shtml