GubaLex: Guba-oriented sentiment lexicon for big texts in finance

Yunchuan Sun IEEE Senior, Mengting Fang, Xinyu Wang, Shizhe Diao

Abstract—Trading in stock market depends mostly on investor's emotions though technical analysis is a viable tool there. In China, Guba is a typical platform for individual investors to share news and opinions on their favorite stocks. The texts posted in Guba by investors involve in richful emotions which can reflect their willingness on the stock. Few works focus on Guba sentiment analysis though numerous have been done on investor sentiment analysis in finance market for the purpose of understanding the market. Text mining is the most popular method to analyze the sentiment implied in the web text, which depends heavily on the lexicon. Existed lexicons for general purpose work badly on sentiment analysis for Guba messages. In this work, we construct a specified lexicon for Chinese Guba, named GubaLex, in considerations of the characteristics of the Guba text: short, emotion enriched, colloquial (informal), and stock market oriented. It is constructed by using the merge of HowNet and NTUSD as the basic sentiment lexicon, then adding stock terms from the Guba corpus and information in the area of stock market. Based on GubaLex, we develop the bullish lexicon GL-Bull and the bearish lexicon GL-Bear especially including bullish and bearish sentiment terms for further sentiment analysis. We also proposed an auto update module and sentiment classification algorithm for Guba texts. The experiments show the proposed lexicon works better in sentiment analysis than the previous, like HowNet and NTUSD.

1 Introduction

As a typical representative of animal spirits, emotion plays a significant role in people's irrational economic decision-makings [1]. With Donald Trump took the lead and finally won the final success in the just-ended US presidential election, many investors got into the panic with sharply rising risk aversion, which thus triggered the gold prices surging over 5%. However, the price soon fell back as people gradually returned calm. Meanwhile, the share prices in global stock markets went down sharply and then recovered. Similar conditions occurred in UK for the Britain left recently. The emotional changes of individual investors directly affect their investment decisions. Therefore, it is no wonder that these black swans can cause global mood swings and thus influence the world economy. What's more, it is no exaggeration to say that almost all of these turmoils in financial markets are caused by investor sentiment.

To date, there have been a lot of studies on the relationships between investor sentiment and financial decisions. Slovic(2000) found that people's perception of risk and benefit is associated with their emotions. The stronger the emotional intensity, the greater the negative correlation between the risks and benefits. Lucey and Dowling(2005) found that the expectation of one stock's future performance mainly depends on the investor's impression of the company rather than rational analysis. Nofsinger(2005) also found that the social emotion can affect people's financial decision-makings through the human interaction in real life. However, the subjective feelings of investors were difficult to measure, and the means and methods of obtaining investor sentiment had been highly limited at the very beginning. Fortunately, with the rapid development of information technology, more and more people get used to expressing and communicating their feelings and opinions through the internet, and the emergences of Web 3.0 [2], mobile cloud computing [3], etc. provides us with unprecedented convenience for sentiment mining and analysis of massive online texts.

In today's global financial markets, China plays an inestimable role as the world's second largest economy, and it is a unique phenomenon that the China stock market is mainly composed of retail investors. Therefore, it is extremely valuable and necessary to exploit the sentiment of Chinese investors. Furthermore, most of these individual investors are easily affected by rumors, following the crowd blindly and chasing sell, thus resulting in heavy losses and market turmoils. In this case, they desperately need a platform to get useful information and exchange investment experience, which brings out Guba — a newly emerging financial social platform unique to China.

Through Guba, stock investors can share their feelings and opinions toward the market by creating, reading or replying posts. By these means, individual investors can

Y. Sun is with the International Institute of Big Data in Finance, Business School, Beijing Normal University, China, Email: yunch@bnu.edu.cn.
M. Fang is with School of Mathematical Science, and the International Institute of Big Data in Finance, Beijing Normal University, China
X. Wang is with School of Mathematical Science, and the International Institute of Big Data in Finance, Beijing Normal University, China
S. Diao is with College of Information Science and Technology, and the International Institute of Big Data in Finance, Beijing Normal University, China

0-7695-6304-X/17/31.00 ©2017 IEEE
DOI 10.1109/SKG.2017.00013

exchange news and opinions, trying to optimize their investment strategies and gain more profits. Investors' emotions have been spreading through massive stock posts in Guba, which continuously affect people's investment attitudes and decisions, thereby influence the whole stock market. Through the mining and analysis in Guba, investor sentiment can be efficiently reflected to regulators, thus enabling them to timely acquire and control the market dynamics so as to guard against systematic financial risks. However, considering the constantly updated stock terms and the special expression way through investors, traditional lexicons and text analysis algorithms are no longer applicable. Thus, it is necessary to build a specialized stock sentiment lexicon which can help us efficiently explore the investor sentiment reflected by stock posts in Guba.

The remainder of the paper is organized as follows. In Section 2, related works are introduced. Section 3 presents the features of Guba texts. In Section 4, we describe the datasets of our work. The details of the construction process of GubaLex are demonstrated in Section 5. Section 6 introduces an auto update module for GubaLex. In Section 7, we propose a sentiment classification algorithm for Guba texts to evaluate the efficiency of GubaLex. Finally, Section 8 concludes the paper and discusses the future work.

2 Related Works

A common definition of emotion in psychology is a complex state of feeling resulting in physical and psychological changes that influence human's thought and behavior. Russell et al. deemed that any problem people faced with contains emotional factors, and financial decision-making is no exception. However, many classical modern financial theories, like the capital asset pricing model(CAPM), exclude human emotion from their research scopes. a great deal of behavioral finance research results indicate that the emotion of investors have remarkable effects on financial decisions, particularly with high risks and uncertainty [4], [5]. With the development of behavioral finance in recent years, financial sectors are paying more and more attention on the investor sentiment, thus boosting the emotion research in financial areas.

As the core of emotion research, sentiment analysis is aim to identify the emotional tendency of a given material. The current research methods are largely based on lexicon, corpus and machine learning. Owning to the features of easy to use and good expansibility, sentiment lexicon is the most essential and widely used tool in sentiment analysis. In the early days, many researchers extended the manually collected sentiment words through existing dictionary resources to obtain a large number of sentiment words. Fellbaum et al. produced WordNet in 1998 inspiring by psycholinguistic theories of human lexical memory, and it has become the most widely used on-line lexical reference system up to now [6] and the foundation of many developed lexicons. Strapparava and Valitutti put forward a a linguistic resource for the lexical representation of affective knowledge evolved from WordNet [7]. Baccianella et al. presented a freely available lexical resource SentiWordNet in which each synset of WordNet is tagged with three labels *Objective*, *Positive* and *Negative* [8]. Hamouda et al. constructed a Machine Learning Based Senti-word Lexicon (MLBSL) outperformed than SentiWordNet using Amazon corpus [9]. Considering the detailed and subtle subjectivity relations between the different participants of a verb, Maks et al. proposed a lexicon model for the description of verbs to be used in deeper sentiment analysis and opinion mining [10]. Ruppenhofer et al. created the FrameNet lexical database as an on-line lexical resource for English on the basis of frame semantics with more than 10,000 lexical units [11].

As the most complicated and changeable language, Chinese has its own unique lexicons for sentiment analysis. Dong et al. built HowNet as a renowned Chinese-English bilingual knowledge base describing relations between concepts and their attributes [12]. On the basis of HowNet, Chen et al. proposed a universal concept representational mechanism called Extended-HowNet by extending the word sense definition mechanism [13]. National Taiwan University organized and released a Chinese sentiment polarity dictionary named NTUSD.

However, most studies are focused on building general-purpose sentiment lexicons applying to any domain, while relatively little work has been done on domain-specific lexicons. Considering that there are many frequently occurred stock terms under specific Guba contexts, like '割肉'(cut meat, which means selling stocks at a much lower price), '套牢'(hung up, which means being trapped in the stock market) and so on, which can be seldom seen in formal language or even will never appear in situations unrelated to stocks, traditional sentiment lexicons are no longer effective. To the best of our knowledge, there is not a well-developed sentiment lexicon designed specifically for the stock market at present, which is such a big defect in financial sentiment analysis. In this paper, a specific lexicon is developed for sentiment analysis for Guba messages, which provides a useful tool for stock-related text analysis and mining.

3 Features of Guba texts

As a financial social networking, Guba is currently one of the most widely used platform for stock information

exchange in China. Through it, millions of individual investors share news, opinions, and experiences in time easily. A specific discussion column is set up for each listed company in Guba. In summarization, Guba text is characterized with the following features.

- **Big volume.** There are nearly 3000 stocks up to now in China A-share market and each of them involves in large numbers of investors, which naturally results in a big volume of stock posts in Guba. Take Sina Guba as an example, the number of stock posts has reached more than 8 million since foundation.

- **Short text.** As a typical kind of small text data, stock posts in Guba have strict space limitations and the theme of each post is even confined to only one sentence. In order to grasp other investors' attention from fast browsing, posters have to simplify the expression, trying to fully convey their emotions in a short theme and post, which makes the stock post data have strong efficiency.

- **Emotion enriched.** The stock posts in Guba tend to contain much more strong emotions rather than ordinary financial texts. Investors can express their emotions by posting or replying Guba messages in real time. For example, investor may post '牛！继续涨啊！'(Great! Keep on rising!) when the stock price continually rises, while '完蛋了！'(It's over!) if the shares go down constantly. These emotion-enriched posts timely reflect the investor's sentiment in Guba, providing precious resources for sentiment analysis.

- **Colloquial(informal).** In Guba, people can freely express their opinions and emotions on the market. The words posted in Guba is much more colloquial than the formal news and official announcements from government, companies, and analysts and economists. Investors usually share news and opinions through Guba messages in a relaxed and casual way. They often use abbreviations and nicknames for simplicity, e.g. '花花' for '同花顺'(Straight flush, with the stock code 300033) and 'vol' for 'volume'. What's more, many vivid and humorous words created by investors are also frequently used in Guba, e.g. '杀跌'(sell into corrections) and '多头'(bull position), which are non-standard or are hard to understand in formal language. These colloquial nicknames are easily used to express their tendentiousness on some stock, and really useful for the sentiment analysis of Guba texts.

- **Stock-Market-Oriented.** Many works have been done on the relationship between the emotions of individuals and the stork market behaviors, and most utilize social media like Twitter [14], Facebook [15], Message Board [16], etc. However, few researchers in this area focus on stock investors. In China, there are more than 80 percents of retailer investors on stock market in China while only 20 percents in USA. Due to the amount of individuals, Guba emerged and is becoming more and more popular. In Guba, nearly all of the users are stock investors, trying to obtain or share stock relevant information by reading or creating posts. Therefore, post data in Guba is specifically oriented towards the stock market, which has much stronger efficiency in our analysis.

- **Strong interactive.** Compared with traditional finance news and blogs, the stock posts in Guba have much stronger interactivity with investors. Through the network, investors can express their feelings and opinions by posting and replying in Guba at any time and any place, which makes the interaction and information propagation among investors much faster and more convenient.

The features mentioned above make the sentiment analysis in Guba much different compared with in other areas, while existing lexicons contain little or even no specialized stork terms and common words used by investors. Thus, a specific lexicon is of great necessity for Guba sentiment analysis.

4 DATASETS

In this work, we build a distributed web crawler to automatically fetch stock posts from Sina Guba and Eastmoney Guba, thus to construct a large Guba corpus. Sina Guba and Eastmoney Guba are the most mainstream Guba platforms with a huge amount of active users currently in China. According to statistics, the stock posts in our Guba corpus span from 2008 until now, containing more than 500 million stock posts corresponding to more than 1.5 million users with specific ID and many more anonymous users. In addition to the Guba corpus we built, we have also collected a great number of stock relevant data, which includes stock-related media news, basic market data and the information of nearly 3000 Chinese A-share listed companies which we have already bought.

5 GUBALEX CONSTRUCTION

As time goes by, the stock market expressions of investors has never stopped changing. China stock market was officially established in 1990, and at that time, the earliest stock terms are very limited. With the rapid development of China stock market and the popularization of internet, numerous novel and

imagery words have been popping up. Confronted with these highly specialized and ever changing stock relevant words, the existing traditional lexicons no longer apply for today's research. In order to execute in-depth excavation and analysis of the investor sentiment in Guba, we need to construct a professional stock sentiment lexicon directed towards Guba. Considering that people's language expression is essentially inseparable from the most basic words and sentence patterns, we first combine two authoritative Chinese lexicons to construct a basic sentiment lexicon. We then add a quantity of stock terms from the Guba corpus and information in the area of stock market into the lexicon. Finally, we extract terms specifically reflecting bullish emotions and bearish emotions to construct lexicons GL-Bull and GL-Bearish respectively.

5.1 Basic Sentiment Lexicon

Although the language expression of stock posts in Guba has its special features, most of them still rely on the basic Chinese expression way. Therefore, we directly choose the union of HowNet and NTUSD as our basic sentiment lexicon. HowNet is a relatively comprehensive and systematic commonsense knowledge base both in China and abroad, while NTUSD is also an authoritative Chinese semantic lexicon developed by National Taiwan University. The Chinese words in the union set can cover almost any word or phrase for normal use. In addition, we eliminate some unfamiliar words manually which would not appear in our daily expressions. Finally, we obtain 47274 basic sentiment words overall.

5.2 Stock Terms

However, most of these special stock words are out of the range of our daily use. This means we can't accurately and comprehensively identify the specialized stock terms only by our basic sentiment lexicon, which brings great difficulty to the analysis and mining of stock-related information. To solve this very problem, we collect 1429 stock terms through the investigation of vast stock related materials, which came from financial news, stock comments, information of Chinese A-share listed companies, and especially our Guba corpus. Owning to the strong interactivity and pertinence of Guba as a highly centralized place for stock investors, we not only harvest a great number of stock jargons, but also obtain each stock's nickname from investors' posts and replies, which are actually used much more frequently than the real name. For example, investors in Guba often call the stock '同花顺'(Straight Flush 300033) as '花花', which is just the second word '同花顺'(Straight Flush 300033). These nicknames usually spring from Chinese homonym or the abbreviation of their real names.

5.3 Bullish and Bearish Sentiment Terms

Stock investor sentiment is usually divided into three categories: bullish, neutral and bearish. Investors usually tend to continue holding their shares or buying more when they are in a bullish mood, while bearish investors are more likely to selling the stocks. When the stock trend remains stable or the market prospect is unclear, investors often hold a neutral attitude, collecting information and watching the situation without buying or selling stocks. As a tool to convey investors' attitudes and opinions, stock posts in Guba can also be classified into three types according to sentiment taxonomy. Through the observation of a large number of posts in Guba, we find that there is typically at least one sentiment word explicitly expressing poster's opinion in a bullish or bearish post. However, neutral posts are mostly objective information about stocks, but not including sentiment words. Since our focus is mainly on specific investor sentiment, we also construct lexicons GL-Bull and GL-Bear containing bullish and bearish words by extraction from the above lexicon we built. In Table 1, we present part of results of the bullish and bearish word list.

TABLE 1: Part of bullish and bearish words in GubaLex

看涨词语(Bullish Words)	看跌词语(Bearish Words)
上涨(rise)	下跌(fall)
回升(rebound)	跳水(dive)
黑马(dark horse)	劣马(inferior horse)
绩优(high quality)	乌云(dark clouds)
解套(get rid of a trap)	套牢(hung up)
龙头股(leading shares)	垃圾股(junk stocks)
阳线(positive line)	阴线(negative line)
坚挺(strong)	崩盘(collapse)

Ultimately, we successfully construct the stock sentiment lexicon oriented Guba called GubaLex, including 47274 basic words and 1429 stock terms. As a part of GubaLex, the bullish and bearish lexicon GL-Bull and GL-Bear contain 6603 and 11338 sentiment words respectively. The final composition of GubaLex is shown in:

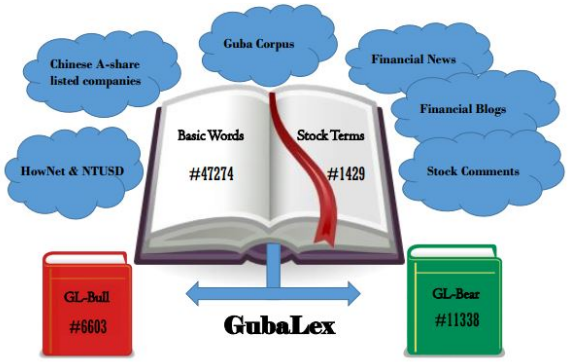

Fig. 1: The composition of GubaLex

6 AUTO UPDATE MODULE

Considering the accelerating update speed of network language, traditional static lexicons have been unable to meet social development needs of the times no longer. Especially in Guba, where large numbers of individual investors communicate and exchange their opinions and experiences every day, new words keep emerging in an endless stream. Although the GubaLex have already contained the great majority of words frequently used by investors, there still exist a certain amount of words we ignored or need to be updated. For example, we have observed that the word '扎心'(sorrowful), which used to be a dialect of Northerners in China, has appeared frequently on the internet in order to describe an extremely sorrowful and distressed feeling starting from this year. And this word also has been widely used in Guba when stock prices drop or investors lose their money. This word has not been collected when we constructed GubaLex, however, it is very important and is necessary to be added in our lexicon.

In this section, we would like to introduce a new-word-identification algorithm and use the Guba corpus to test and validate the effectiveness of our algorithm.

6.1 A New-Word-Identification Algorithm

As we know, spaces which used to separate words, numbers and punctuation, play an important role as words boundary in English writing, can help computers recognize words without ambiguousness. However, in Chinese writing, words adjacent in the same sentences are not separated by space. As a consequence of inexistence of space, computer cannot recognize words correctly, not to mention understanding sentences. So, it is necessary to recognize natural-language words before we do any follow-up works.

We build an auto-update modular to recognize new words based on PAT-Array and information entropy. PAT-Array is a retrieval algorithm developed by Manber and Myers in 1993 [17], it has been used to obtain the boundary of English phrase in a large text and achieve a great performance [18], [19]. We can regard every single Chinese character as an English word, so we can use PAT-Array to recognition Chinese words. To improve the accuracy of new words recognition, we also calculate the information entropy of prefix I_p and suffix I_s of words found by PAT-Array. Information entropy is also used to recognize new words and phase in text. In a word, the diversity of the prefix and suffix of natural-language words should be rich, while that of meaningless phrase would be poor. For instance, the diversity of the prefix of '盘手'[a fragment of '操盘手' (trader)], follows the word '操' in the vast majority of cases, should be poor. We find all the n-words phrase appears in the Guba, and then calculate the information entropy of prefix and suffix of them to quantify the diversity, which is defined as

$$L(\omega) = \frac{1}{n} \sum_{a_i \in a} c(a_i, \omega) log(\frac{c(a_i, \omega)}{n}) \quad (1)$$

where a is the set of prefix or suffix of ω, and $c(a_i, \omega)$ is the frequency of a_i is adjacent to ω [20]. If an n-words phrase were whole, both its prefixal and suffixal information entropy should be large, whereas either of them might be low.

The update algorithm is shown in the following:

Algorithm 1 Auto Update Module

Input: input Text
Output: output Words

1: Grab texts from Guba
2: Remove non-Chinese character and words included in GubaLex
3: Search n-words phrase
4: Calculate I_P and I_S
5: **if** $I_P > threshold$ and $I_S > threshold$ **then**
6: **return** phrase
7: **end if**

6.2 Experiments

To test the ability and precision of the auto update module of our model, we use the update algorithm to search the entire Guba corpus we built and then identified the new words via our model. Finally, we have found 578 candidate words and 472 words of them are certainly new words through our manual identification with the precision 81.7%. Meanwhile, we select sentiment words from the new-found words and classified them into bullish and bearish. In Table 2, we present part of results of the new words.

TABLE 2: Part of new words in Guba

看涨词语(Bullish Words)	看跌词语(Bearish Words)
股神(master of stock)	破位(breaking down)
抗跌(resist the crash)	唱空(sell-off)
给力(awesome)	狗庄(damn market)
高开(high opening)	破净(breaking the net)
金股(golden share)	韭菜地(small retail investors)

These words are certainly new words which are frequently used by individual investors in Guba, but haven't been collected in our prior work. Through the auto update module, we complement and improve the original lexicon by adding 472 words, including 258 bullish words and 214 bearish words. Moreover, according to the efficiency of our auto update module, we plan to update the lexicon regularly in order to keep enriching and perfecting GubaLex.

7 EVALUATION

In this section, we use Guba corpus that we built in the preceding section to test the validation of our lexicon. Firstly, we do data preprocessing of Guba corpus by stock posts filtering to eliminate the noise for further analysis. We randomly selected 500 posts from the corpus in recent three years for testing and manually annotate each post according to the investor sentiment as *bullish*, *bearish* or *neutral*. We then develop a sentiment score computation algorithm and compare the program results using HowNet, NTUSD and GubaLex with the manual annotation so as to verify the efficiency of GubaLex in sentiment classification.

7.1 Stock Posts Filtering

As mentioned before, the stock posts data in Guba is enormous and contain multiple types of information. On the one hand, it will take several hours to be processed if used as it is. On the other hand, we only care about the sentiment of individual investors in Guba, rather than news or announcements posted by some official accounts, which can cause interference to the sentiment analysis. Therefore, we filtered out the irrelevant posts and reserved only those posts which are more likely to express the feeling of investors in Guba.

7.2 Sentiment Score Computation

We first segment the words of each sample post (SP) using JIEBA, a mature Chinese word segmentation tool, to provide the basis for Chinese language processing. We denote the result of segmentation by $SP = \{w_1, w_2, \cdots, w_n\}$. We choose the bullish and bearish sentiment words in each SP as stock-emotional features $Bull_{SP} = \{blw_1, blw_2, \cdots, blw_l\}$ and $Bear_{SP} = \{brw_1, brw_2, \cdots, brw_m\}$ based on GL-Bull and GL-Bear, where $Bull_{SP}$ and $Bear_{SP}$ are subsets of SP, and $blw_i \in$ GL-Bull, $brw_j \in$ GL-Bear.

In order to avoid the superposition and confusion of bullish emotions and bearish emotions, we will calculate the bullish sentiment score and bearish sentiment score using GL-Bull and GL-Bear separately. Then we use the following sentiment analysis algorithm to get the stock-emotional tendency of each sample post.

- **Sentiment words.** Sentiment words refer to the words that are used to express people's emotions. In this work, the focus is especially concentrated on the bullish sentiment and the bearish sentiment of investors. According to the sub-lexicons GL-Bull and GL-Bear specialized for bullish and bearish sentiment terms, we first search and locate the sentiment words in each sample post. If a bullish word is retrieved, the bullish sentiment score add 1, whereas the bearish sentiment score add 1 when a bearish word appears.

- **Degree words.** Degree words are used to qualify or modify sentiment words in order to make the expression more precise and subtle. For instance, an investor posting *sharp rise* tends to have more confidence in the stock prices rising than the investor who uses *slight rise*. When a sentiment word is located, we need to search forward for degree words and give corresponding weights for different degrees. According to the intensity of degree words, we assign the highest degree words like "extremely", "seriously" with the weight 4; the second-highest degree words like "relatively", "fairly" with the weight 2; the lowest degree words like "slightly", "merely" with the weight 0.5. The original score of each sentiment word with a nearby degree word will be multiplied by its corresponding weight so as to obtain the final score.

- **Exclamation mark.** Exclamation mark denotes powerful and intense emotion. When an exclamation mark appears in a post, we will multiply the original score times two.

- **Negative words.** Negative words express negative meanings in a sentence, such as "no", "never", "hardly" and so on. Considering the complex forms of double negative and multiple negatives, we not only have to locate the negative word, but also need to count the occurrences of negative words. If the occurrence number is singular, we then add the corresponding sentiment score to the other side; if it is an even number, the score can remain intact.

According to the above processing principle, we get the score of each sentiment word. We then sum up the score of all the bullish words and bearish words separately in a stock post as its bullish score $SBull(SP)$ and bearish score $SBear(SP)$, and thus get the total score $TS(SP)$ as follows:

$$TS(SP) = SBull(SP) - SBear(SP) \tag{2}$$

where

$$SBull(SP) = \sum blw_i \tag{3}$$

$$SBear(SP) = \sum brw_j \tag{4}$$

7.3 Score Mapping

For simplicity, we use three values 1, 0, -1 to represent the bullish, neutral and bearish sentiment respectively. We map the value to the three sentiment using the following rules.

$$Polarity = \begin{cases} 1 & \sum TS(SP) > 0 \\ 0 & \sum TS(SP) = 0 \\ -1 & \sum TS(SP) < 0 \end{cases} \quad (5)$$

7.4 Results

For comparison, we conduct the sentiment classification of sample stock posts using HowNet, NTUSD and GubaLex respectively.

The outputs of using different lexicons can be evaluated by:

$$Precision = avg(Prec(bl) + Prec(br) + Prec(neu)) \quad (6)$$

$$Prec(bul) = \frac{N_{cor}(bl)}{N_{all}(br)} \quad (7)$$

Where Precision is the average precision of our system generated results in different classes (bullish class, bearish class, neutral class). $N_{cor}(bl)$ is the amount of correct bullish posts compared with manually labeled results. $N_{all}(bl)$ is the number of the manually annotated bullish documents in the evaluation texts. The results of the experiment are reported in Table 3.

TABLE 3: Sentiment classification result using HowNet,NTUSD,GubaLex

	Prec(bl)	Prec(br)	Prec(neu)	Precision
HowNet	0.224	0.218	0.950	0.516
NTUSD	0.269	0.232	**0.985**	0.482
SEL-Guba	**0.833**	**0.768**	0.906	**0.844**

We summarize the results in Table 2 and highlight in bold font the best performance under each lexicon. As a whole, the proposed lexicon GubaLex is quite efficient to classify the sentiment tendency of stock posts in Guba. The improvement is mainly due to the increasing in the stock-related terms and the specialized bullish and bearish sentiment terms. Compared with manually annotated result, the total precision using GubaLex is 84.4%, which is significantly higher than using HowNet(51.6%) and NTUSD(48.2%). The good results of this experiment are mainly due to the following reasons. One is that, we successfully construct a Guba oriented stock sentiment lexicon, GubaLex. It can effectively perform basic analysis (e.g. word segmentation) on Guba messages, thus providing a good foundation for further research. Another reason is that the sub sentiment lexicons GL-Bull and GL-Bear play a significant role in accurately identify the emotional tendency of Guba messages.

Figure 2 shows the results for the sentiment classification using HowNet, NTUSD and GubaLex. Compared to HowNet and NTUSD, GubaLex gains much higher precision in identifying bullish posts and bearish posts. For neutral posts, the tree lexicons have similar performance, since this kind of posts mainly contains news and announcements about one stock instead of clear bullish or bearish attitudes of investors, which is not the focal point in this work.

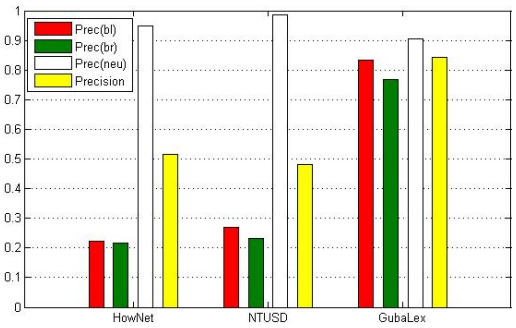

Fig. 2: Precision using different lexicons

8 Conclusion and Future Work

In this paper, we construct a Guba-oriented stock sentiment lexicon GubaLex to be used in deeper sentiment analysis and opinion mining of stock-related information. And we have demonstrated that the lexicon we built using a sentiment analysis algorithm developed on the basis of GubaLex could achieve a better result in sentiment classification tasks of stock posts than existing lexicons. We have also introduced an auto update module in order to regularly enrich our lexicon, and the experimental results have demonstrated the feasibility of our method.

With the era big data finance approaching, more and more stock-related data are generated in substantial amounts all the time, which can be efficiently used in the construction of our lexicon. As future work, we will extend our approach to multiple financial resources and apply GubaLex into more practical areas, such as stock predication and investment recommendation.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant NO. 61371185.

References

[1] B. M. Lucey and M. Dowling, "The role of feelings in investor decision-making," *Journal of Economic Surveys*, vol. 19, no. 2, pp. 211–237, 2005.

[2] B. P. Torres and A. G. González, *Evolution of the Semantic Web Towards the Intelligent Web: From Conceptualization to Personalization of Contents*. Springer International Publishing, 2017.

[3] G. Skourletopoulos, C. X. Mavromoustakis, G. Mastorakis, J. M. Batalla, C. Dobre, S. Panagiotakis, and E. Pallis, *Towards Mobile Cloud Computing in 5G Mobile Networks: Applications, Big Data Services and Future Opportunities*. Springer International Publishing, 2017.
[4] J. P. Forgas, "Mood and judgment: the affect infusion model (aim)." *Psychological Bulletin*, vol. 117, no. 1, pp. 39–66, 1995.
[5] G. F. Loewenstein, E. U. Weber, C. K. Hsee, and N. Welch, "Risk as feelings." *Psychological Bulletin*, vol. 127, no. 2, 2001.
[6] C. Fellbaum and G. Miller, "Wordnet : an electronic lexical database," *Cognition Brain & Behavior*, 1998.
[7] R. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *International Conference on Language Resources & Evaluation*, 2004, pp. 1083–1086.
[8] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta*, 2010, pp. 83–90.
[9] A. Hamouda, M. Marei, and M. Rohaim, "Building machine learning based senti-word lexicon for sentiment analysis," *Journal of Advances in Information Technology*, vol. 2, no. 4, pp. 31–34, 2011.
[10] I. Maks and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications," *Decision Support Systems*, vol. 53, no. 4, pp. 680–688, 2012.
[11] J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, and J. Scheffczyk, "Framenet ii: Extended theory and practice," in *World Wide Web Conference Series*, 2010.
[12] Z. Dong and Q. Dong, "Hownet - a hybrid language and knowledge resource," in *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings*, 2003, pp. 820–824.
[13] K. J. Chen, S. L. Huang, Y. Y. Shih, and Y. J. Chen, "Extended-hownet- a representational framework for concepts," 2005.
[14] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, p. 1 – 8, 2010.
[15] Y. Karabulut, "Can facebook predict stock market activity?" *Ssrn Electronic Journal*, 2011.
[16] T. H. Nguyen, K. Shirai, and J. Velcin, *Sentiment analysis on social media for stock movement prediction*. Pergamon Press, Inc., 2015.
[17] U. Manber and G. Myers, "Suffix arrays: a new method for on-line string searches," in *Acm-Siam Symposium on Discrete Algorithms*, 1990, pp. 319–327.
[18] M. Zhou and F. W. Tompa, *The suffix-signature method for searching for phrases in text*. Elsevier Science Ltd., 1998.
[19] M. GALLé, P. PETERLONGO, and F. COSTE, "In-place update of suffix array while recoding words," *International Journal of Foundations of Computer Science*, vol. 20, no. 6, pp. 1025–1045, 2012.
[20] Q. Ma and F. Xia, "Proceedings of the second sighan workshop on chinese language processing - volume 17," in *Sighan Workshop on Chinese Language Processing*, 2003, pp. 215–21.