





# A Novel Way to Build Stock Market Sentiment Lexicon

Yangcheng Liu<sup>1</sup>  and Fawaz E. Alsaadi<sup>2</sup> 

<sup>1</sup> School of Business Administration, Southwestern University of Finance and Economics, No. 555, Liutai Avenue, Wenjiang Zone, Chengdu 611130, China  
liuyangcheng@outlook.com

<sup>2</sup> Department of Information Technology,  
King Abdulaziz University, Jeddah, Saudi Arabia

**Abstract.** The construction of domain-specific sentiment lexicon has become an important direction to improve the performance of sentiment analysis in recent years. As one of the important application areas of sentiment analysis, the stock market also has some related researches. However, when considering the heterogeneity of the stock market relative to other fields, these studies ignore the heterogeneity of the stock market under different market conditions. At the same time, the annotated corpus is also indispensable for these studies, but the annotated corpus, especially the social media corpus that is not standardized, domain-specific and large in volume, is very difficult to obtain, manually labeling or automatic labeling has certain limitations. Besides, in the evaluation of the stock market sentiment lexicon, it is still based on the general classification algorithm evaluation criteria, but ignores the final application purpose of the sentiment analysis in the stock market: helping the stock market participants make investment decisions, that is, to achieve the highest profit. To address those problems, this paper proposes an unsupervised new method of constructing the stock market sentiment lexicon which based on the heterogeneity of the stock market, and an evaluation method of stock market sentiment lexicon. Subsequently, we selected four commonly used Chinese sentiment dictionaries as benchmark lexicons, and verified the method with an unlabeled Eastmoney stock posting corpus containing 15,733,552 posts about 2400 Chinese A-share listed companies. Finally, under our lexicon evaluation framework which based on the portfolio annualized return, the stock market sentiment lexicon constructed in this paper has achieved the best performance.

**Keywords:** Sentiment lexicon · Stock market · Investor sentiment

## 1 Introduction

In recent years, with the rise of social media (such as Weibo, WeChat and some forums), posting has become one of the most popular behaviors in the Internet age. These user-generated content makes social media the largest source of data for opinion mining. However, social media data has the characteristics of fast generation and large volume. It is impossible to manually analyze social media data. Therefore, some methods such as sentiment analysis that automatically mine large amount of opinion

data have been used in those days. Some existing sentiment analysis methods usually use supervised learning algorithms such as support vector machine [8], naive Bayesian method [17], fuzzy TOPSIS [9], integrated learning [21] and other deep learning method [11], but these algorithms require labeled data as support. Although this method has higher accuracy when the training data and the testing data come from the same domain, the labeled data, especially the labeled social media data is difficult to obtain, and manual labeling is also time consuming and labor intensive. Therefore, the use of sentiment lexicon for sentiment analysis is an important direction in the field of sentiment analysis.

For the stock market, there have been some studies on the use of sentiment analysis to predict stock market related variables [1, 10, 18, 19, 25], such as stock prices and trading volumes. The use of social media data for stock market related decision-making research has shown an upward trend in recent years, which is closely related to the frequent activities of users on social media. At the same time, the acquisition of social media data has its incomparable advantages in terms of difficulty and timeliness when compares to traditional data. These advantages help stock market participants to evaluate the stock market in real time, which is invaluable for stock market investment decisions during the trading day. Some existing researches mainly use the existing general sentiment dictionaries to analyze the stock market investor sentiment, but many previous studies have shown that sentiment words will change with different application fields and contexts [7, 15].

However, there are few studies on the construction of sentiment lexicon in the stock market. Loughran and McDonald [10] used the official text data which comes from the American Securities and Exchange Association from 1994 to 2008 to manually construct a financial sentiment lexicon. Mao [12] used the current stock return rate to label the news corpus, and proposed method to automatically construct the Chinese financial sentiment lexicon. Nuno [14] improved the existing method by comparing three kinds of statistical based methods to construct sentiment lexicon, and proposed two supplementary statistical indicators to improve the existing method.

Although the above study considers the heterogeneity of the stock market relative to other fields, it ignores the heterogeneity of the stock market under different market conditions. For example, consider two situations: (1) the market fell, a stock rose against the trend; (2) the market rose, a stock rose. If a single stock market sentiment lexicon is used for sentiment analysis, in the above two cases, the related sentiment words have the same sentiment intensity. However, considering the real word investment decision-making, in these two cases, the investment decision is completely different, and the stock that rises against the trend should be the preferred investment target of the stock market participants. At the same time, although the sentiment analysis method based on sentiment lexicon itself is an unsupervised learning method, it is a supervised method for the stock market sentiment lexicon construction method itself, and also requires an annotated corpus. However, previous studies have shown that sentiment analysis based on supervised algorithms is a better choice for labeled data. In addition, in the evaluation of the stock market sentiment lexicon, the general classification criteria is still used as the main method, but the final purpose of the sentiment analysis applied to the stock market is neglected: to help the stock market participants make investment decisions, that is, to achieve the highest profit.

Therefore, in order to solve the above problems, this paper proposes a new unsupervised stock market sentiment lexicon construction method. The main contributions of this paper are: (1) According to the heterogeneity of the stock market in different market conditions, the sentiment seed words are automatically extracted, and the bullish sentiment lexicon and the bearish market sentiment lexicon are constructed based on the automatically extracted seed words. This method not only helps to resolve the heterogeneity of the stock market relative to other fields, but also considers the heterogeneity of the stock market itself. (2) An unsupervised method for automatically constructing stock market sentiment lexicon is proposed, which eliminates the need of labeled corpora (3) transforms sentiment lexicon optimization goals, and aims at the realistic goal of sentiment analysis in the stock market: help Stock market participants make investment decisions.

## 2 Related Works

With the advent of the Internet era, data acquisition has become more and more convenient. The stock market has been a hot topic in the research field for many years, more and more research uses the large amount of data obtained in the Internet era to study the stock market. Nayak [13] proposes a novel condensed polynomial neural network (CPNN) for the task of forecasting stock closing price indices. Rashid [16] uses a large panel of Pakistani non-financial firms over the period 2000–2013 to examine the role of financial Constraint in establishing the relationship between cash flow and external financing. Challa [3] try to through light on investment decisions by linking it with beta values of respective stocks. But most of the above studies are based on the effective market, and the irrational behavior of investors, for example: investor sentiment [22], the herding behavior [26], is not considered.

The construction of the stock market sentiment lexicon is a domain-specific task, Loughran and McDonald [10] used to form a financial sentiment lexicon manually with the official text data from the American Securities and Exchange Association from 1994 to 2008. Mao [12] proposed a method based on Chinese corpus to construct an sentiment lexicon in the financial field. According to the rise and fall of individual stocks on the day, the news of were classified as “negative” and “positive”, and then the labeled news were used. The news data constructs the financial domain sentiment lexicon by calculating the Jaccard similarity between the seed word and the target word. Oliveira [14] proposed three methods which based on the commonly used methods of PMI, TF-IDF and IG for constructing stock market sentiment lexicon. Finally, it was found that the PMI- based sentiment lexicon construction method has the highest accuracy. Sun [23] built the basic sentiment lexicon with HowNet and NTUSD, then added the unique stock market sentiment word to it, and extended the stock market sentiment lexicon based on information entropy.

In summary, although there have been some studies on the construction of sentiment lexicon in the specific field of the stock market, there are still some shortcomings: First, although the heterogeneity of the stock market relative to other fields is considered, it is ignored the heterogeneity of the stock market under different market conditions. Secondly, although the sentiment analysis method based on sentiment

lexicon itself is an unsupervised learning method, it is a supervised method for the stock market sentiment lexicon construction method itself, and also requires an annotated corpus. However, previous studies have shown that sentiment analysis based on supervised algorithms is a better choice for labeled data. Finally, the evaluation of the stock market sentiment lexicon is still based on the traditional classification evaluation method, ignoring the ultimate goal of sentiment analysis of the stock market: to help stock market participants make investment decisions.

### 3 Methodology

The method of constructing the stock market sentiment lexicon proposed in this paper is mainly composed of three steps: (1) candidate word extraction (2) seed word selection; (3) lexicon expansion. The main idea of the method is to divided the whole corpus into a bullish corpus and a bearish corpus according to the market return. By filtering the sentiment words with the same and stable sentiment in two corpora as seed words, then the seed words is used to expand the sentiment lexicon in the bullish corpus and bearish market corpus respectively. The method proposed in this paper is shown in Fig. 1.

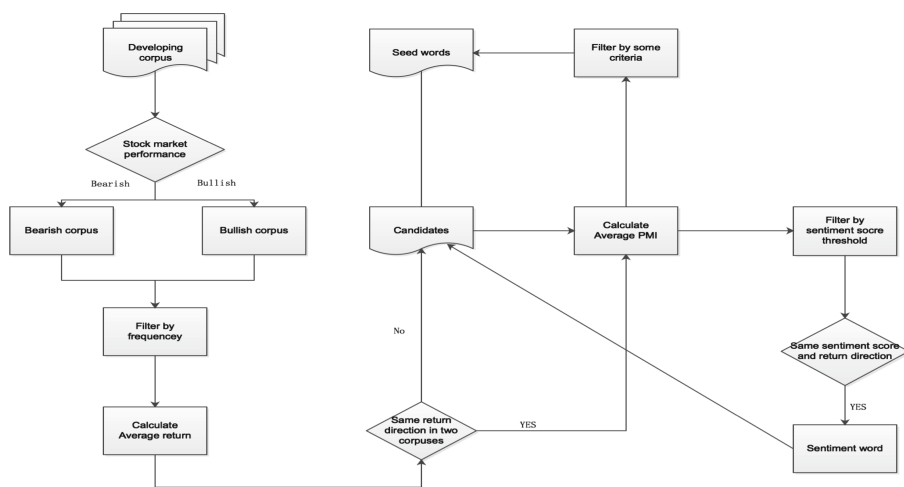


Fig. 1. The framework of the stock market sentiment lexicon construction method

#### 3.1 Data Collection and Processing

This article contains 15,733,552 pieces of posts about 2400 Chinese A-share listed companies. In this paper, 2400 A-share Chinese listed companies has been divided to training data and testing data, to be specific, we use posts data about 70% stocks as the training data, and the remaining posts about 30% stocks as a testing data. The title of all posts constitutes the above training data and testing data. The reason why the post title

is selected as the final corpus is because the text posted in the forum usually shows a more obvious sentiment polarity in the title. The more complex expression in the text will bring too much noise to the final sentiment analysis. In addition, the training set and the testing set are divided in such a way as to ensure the integrity of the daily posting data of each stock, so as to avoid the 70%/30% division of the entire corpus leads to the error of the sentiment analysis of a single stock. The final training set and testing set contain 10,907,590 and 4,371,729 post titles respectively.

Stock-related data comes from the CSMAR database, which mainly includes market daily return of sub-markets, stock daily return. By integrating the corpus with the stock data, this paper uses the training set to construct the stock market sentiment lexicon through the method proposed in this paper, and then uses the testing set to make the final evaluation.

### 3.2 Candidate Extraction

The first step in constructing a stock market sentiment lexicon is to extract candidate word from the corpus of the training set, because not all vocabulary appearing in the corpus is suitable as an sentiment word. The first step is based on Word frequency  $F(I)$  filtering. Although in social media, uncommon expressions often appear, some are not even included in the lexicon, but some uncommon expressions are accepted and widely used in a particular field, such as in stock forums. There are some uncommon expressions such as “raising limit”, but they are very common in the stock market. So if some words are not common, but the word frequency reaches a certain threshold, this article still retains it in the candidate word set.

### 3.3 Seed Word Selection

In previous studies on the construction of sentiment lexicons, the selection of seed words was basically done by manual selection [24, 27] or by automatically labelled data [12]. Different from the previous research, the method of constructing the sentiment lexicon proposed in this paper can automatically extract the seed words from the unlabeled corpus, which not only saves the trouble of labeling the corpus, but also guarantees the extracted seed words are highly field related which contributing to the subsequent expansion of the sentiment lexicon. In previous studies, such as investor sentiment and stock pricing research in behavioral finance [2, 5, 6, 20], the researchers found that the sentiment of stock market participants is closely related to the rise and fall of stock prices. Therefore, this paper believes that the rise and fall of stock prices is often accompanied by the rise and fall of investor sentiment. This assumption has also been supported in previous studies [4].

Drawing on Mao’s [12] study of the economic significance of sentiment words, in this paper, the economic significance of sentiment words is defined as follows:

$$EV(I) = \frac{\sum_{i=1}^n r_i}{n} \quad (1)$$

where  $n$  indicates that the number of posts which contain word  $I$  in the training set, and each of which is associated with a particular stock  $i$ ,  $r_i$  indicates its stock daily return. Simply put,  $E(I)$  is the average stock daily return, for the convenience of calculation, this article uses the base point to represent  $E(I)$ .

With the above definition of the economic significance of sentiment words, the method of automatically extract seed words proposed in this paper has a basis. As previously stated, the rise and fall of stock prices is often accompanied by the ups and downs of investor sentiment, while most stocks should show an upward trend when the stock market is better, and vice versa. Therefore, when the stock market is better, most of the posts in the stock market forum should present more optimistic emotions, so the words'  $E(I)$  appearing in these posts should be greater than zero, and vice versa. And some words that express a stable sentiment polarity should also have a stable economic significance in both cases, that is, regardless of the market is good or bad, these words'  $E(I)$  are the same sign. And these words should be the best choice as a seed word.

At the same time, the sentiment scores calculated by the seed words under different market conditions should also be the same. In this article, we have selected the often used statistic-based algorithm for calculating sentiment scores, the statistic expression is as follows:

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)} \tag{2}$$

Based on above statistics, the sentiment score of a candidate word is calculated as follows:

$$SO_I = \frac{1}{N_{s_{pos}}} \sum_{i=1}^{N_{s_{pos}}} PMI(I, s_{pos}) - \frac{1}{N_{s_{neg}}} \sum_{i=1}^{N_{s_{pos}}} PMI(I, s_{neg}) \tag{3}$$

among them  $I$  is a candidate word,  $S_{pos}$ ,  $s_{neg}$  are positive seed words and negative seed words respectively.  $N_{s_{pos}}$  and  $N_{s_{neg}}$  represent the total number of positive seed words and negative seed words respectively, and in order to ensure the expansion of the subsequent sentiment lexicon is not biased to the difference between the total number of positive and negative seed words,  $N_{s_{pos}}$  and  $N_{s_{neg}}$  must be equal.

Based on the above analysis, the steps to automatically extract seed words from the training set are as follows:

1. Dividing the training set into a bullish corpus and a bearish corpus according to market daily return  $MT$  and calculating candidate words  $I$ '  $EV_{I,Bullish}$ ,  $EV_{I,Bearish}$ ;
2.  $I \in \begin{cases} S_{pos} & \frac{EV_{I,Bearish}}{EV_{I,Bullish}} > 0 \text{ and } EV_{I,Bearish} + EV_{I,Bullish} > 0 \\ S_{neg} & \frac{EV_{I,Bearish}}{EV_{I,Bullish}} > 0 \text{ and } EV_{I,Bearish} + EV_{I,Bullish} < 0 \end{cases}$ , that is, seed words should have stable economic significance in different market conditions;
3. Calculate candidate word seeds  $S_{pos}$ ,  $S_{neg}$  sentiment scores in the bearish market corpus and the bullish corpus:  $SO_{S_{pos,Bearish}}$ ,  $SO_{S_{pos,Bullish}}$ ,  $SO_{S_{neg,Bearish}}$ ,  $SO_{S_{neg,Bullish}}$ ;

$$4. S \in \begin{cases} s_{pos} & \frac{SO_{s_{pos},Bearish}}{SO_{s_{pos},Bullish}} > 0 \quad \text{and} \quad SO_{s_{pos},Bearish} + SO_{s_{pos},Bullish} > 0 \\ s_{neg} & \frac{SO_{s_{neg},Bearish}}{SO_{s_{neg},Bullish}} > 0 \quad \text{and} \quad SO_{s_{neg},Bearish} + SO_{s_{neg},Bullish} < 0 \end{cases}, \text{ that is, seed words should have a stable sentiment polarity in different market conditions.}$$

Although the above 5 steps have been used to extract some seed words with stable economic significance and sentiment polarity in different markets, there are some neutral or misclassified seed words. In order to reduce the errors and noises in the seed concentration, based on the filtering rules designed by Mao [12], this paper designs the following filtering rules:

First of all,  $D(I)$  is defined as all posts which contains candidate word  $I$ , the size of the  $D(I)$  is  $F(I)$ , indicating the number of posts that appear in the corresponding corpus, which is the word frequency. Stock coverage  $SC(I)$  represents the number of stocks which the subdataset  $D(I)$  includes, time coverage  $TC(I)$  represents the number of dates which the subdataset  $D(I)$  included. For candidate seed words, it should have such a feature vector  $\{EV(s), F(s), SC(s), TC(s)\}$ . The specific filtering rules are as follows:

1.  $SC(s) > 200$ : Seed words need to cover more stocks;
2.  $TC(s) > 20$ : Seed words need to cover more dates;
3.  $|EV(s)| > 10$ : Seed words need to have strong economic significance;

After passing the above filtering rules, the paper selected positive and negative seed words respectively. The top 30 seed words are used as the final seed word set. Tables 1 and 2 show some selected the positive seed words and negative seed words that are automatically extracted by the above methods, respectively.

**Table 1.** Positive seed words

$S_{neg}$	$EV_{s,Bearish}$	$EV_{s,Bullish}$	$TC(s)$	$SC(s)$	$F(s)$
涨停	77.12	207.94	244	1538	273324
拉高	36.32	140.46	244	1538	53242
突破	11.61	119.91	244	1537	45898
涨停板	114.57	238.58	244	1535	38917
板	143.37	300.12	244	1520	30213
新区	21.42	193.24	205	1494	25447
献花	3.03	133.26	244	1506	24595
追	45.02	168.59	244	1525	24468
追高	34.00	156.02	244	1517	22852
高开	17.47	127.20	244	1511	22226

### 3.4 Lexicon Expansion

The third step in constructing the stock market sentiment lexicon proposed in this paper is the lexicon expansion, that is, using the seed word set selected above, and calculating the sentiment score of the candidate words according to formula (3). As mentioned

**Table 2.** Negative seed words

$S_{neg}$	$EV_{S,Bearish}$	$EV_{S,Bullish}$	$TC(s)$	$SC(s)$	$F(s)$
跌	-198.05	-34.27	244	1538	376300
跌停	-387.16	-179.45	244	1538	172365
死	-163.93	-1.84	244	1538	117707
砸	-172.60	-24.67	244	1537	117055
下跌	-172.33	-19.57	244	1538	93188
大跌	-162.65	-16.29	244	1537	72860
割肉	-219.41	-37.91	244	1537	69797
减持	-142.99	-1.55	244	1537	65341
垃圾股	-151.44	-3.04	244	1535	57252
破	-189.56	-23.07	244	1535	56802

earlier, the existing research considers the heterogeneity of the stock market field relative to other fields, but ignores the heterogeneity of the stock market itself under different market conditions. This heterogeneity is crucial for making stock market investment decisions through sentiment analysis, because the same rise and fall is completely different for investors in bullish and bearish markets. At the same time, this heterogeneity may have a greater impact on the sentiment polarity and sentiment intensity of some sentiment words. For example, the previously selected neutral word “buy” has the opposite economic significance in the bullish market and the bearish market. This situation may arise because stock market participants are affected by stock market conditions. These neutral words often appear in the headers of affirmative contexts in the bull market, and in negative context posts in bearish markets.

Based on the above considerations, when constructing the stock market sentiment lexicon, this paper divides the corpus of the training set into the bullish corpus and the bearish corpus according to the market return, and constructs the bullish market stock market sentiment lexicon and the bearish market stock market sentiment lexicon respectively.

As with the filtering of seed words, in order to reduce the noise of the final constructed stock market sentiment lexicon and improve its quality, the candidate word in the bullish sentiment lexicon and the bearish market sentiment need to be filtered by the following rules:

1.  $|SO(I)| > SS$ : Candidate's sentiment score needs to be greater than a certain threshold in order to reduce the noise of the final lexicon.
2.  $SC(I) > 30$ : sentiment words need to cover more stocks;
3.  $TC(I) > 3$ : Seed words need to cover more dates;
4.  $|EV(I)| > 1$ : Seed words need to have strong economic significance;



## 4 Evaluation and Results

### 4.1 Evaluation

In the previous study, there is no difference between the evaluation of the stock market sentiment lexicon and the evaluation of the general classification algorithm which evaluates the performance of the sentiment lexicon through traditional classification evaluation criteria such as accuracy rate, recall rate and F-value. But this test method ignores the ultimate goal of applying sentiment analysis to the stock market—helping market participants make investment decisions and maximize investment returns.

Therefore, based on the above considerations, in this paper, we use the bullish stock market sentiment lexicon and the bearish stock market sentiment lexicon constructed in this paper to construct an investment portfolio based on sentiment analysis, and calculate the final portfolio annualized return. The final evaluation criteria in this paper is a comprehensive criteria which combined recall rate and the final portfolio annualized return. At the same time, we selected four commonly used Chinese sentiment dictionaries as the benchmark lexicon, and carried out the same evaluation process, so as to compare the methods proposed in this paper. These four Chinese sentiment lexicons are Hownet, NTUSD, TSING, and DUTIR.

### 4.2 Results

Table 3 shows the final results. The results show that the performance of the sentiment lexicon constructed by the proposed method surpasses the selected benchmark lexicon in terms of the annualized return or the recall rate. In addition, from the results of the benchmark lexicon and the results of the lexicon constructed in this paper, the research on constructing domain lexicon is verified in the previous research. In the face of sentiment analysis in special fields, the general sentiment lexicon does have insufficient sentiment words. The problem of inaccurate sentiment polarity, and the inaccuracy of sentiment is the main problem. Because the sentiment lexicon of the stock market constructed in this paper is not the most in the number of sentiment words, especially the number of negative sentiment words is the smallest, which indicates that the sentiment polarity of the sentiment words of the general sentiment lexicon is not accurate in the special field.

**Table 3.** Main results

Lexicon	Positive words	Negative words	Portfolio return	Recall
Hownet	4528	4320	27.62%	29.90%
NTUSD	2648	7742	71.91%	30.12%
Tsinghua	5567	4469	98.71%	29.09%
no_name	14056	9299	102.38%	63.97%
DUTIR	11205	10763	64.56%	20.65%
Combine	27926	26594	84.09%	63.76%
Our method	6669	4042	<b>640.70%</b>	<b>75.21%</b>

## 5 Conclusion

In the previous research on the stock market sentiment lexicon, although some studies considered the heterogeneity of the stock market relative to other fields, they ignored the heterogeneity of the stock market under different market conditions. At the same time, the previous research on the stock market sentiment lexicon still relies on the labeled corpus when constructing the sentiment lexicon. However, the labeled corpus, especially the social media corpus, is difficult to obtain. There are limitations to automatic labeling according to return, etc. In addition, in the evaluation of the stock market sentiment lexicon, the general classification algorithm is still used as the main method, but the final purpose of the sentiment analysis applied to the stock market is neglected: to help the stock market participants make investment decisions, that is, to achieve the highest profit. In response to the above problems, in this paper, we propose a new unsupervised method for constructing a stock market sentiment lexicon. Through comparison with the four commonly used Chinese sentiment lexicons and a series of tests, we can find that the method of constructing the stock market sentiment lexicon proposed in this paper has been significantly improved.

The research in this paper has certain contributions in both theory and practice. First of all, the method proposed in this paper does not require the labeled corpus as a support, and is an unsupervised method, which can save a lot of manual labor. Secondly, the method proposed in this paper can not only be used to automatically extract seed word sets through the consideration of heterogeneity under different market conditions, but also help stock market participants to make more reasonable investment decisions under different market conditions.

In theory, the change proposed in this paper for evaluating the stock market sentiment lexicon goal is more in line with the reality, and provides a new idea for the evaluation of sentiment lexicon in the absence of labeled data. Secondly, the research on the dynamic relationship between investor sentiment and stock price provides empirical support for related research.

This study still has some limitations, first of all we only use the Eastmoney Guba post which about 2400 Chinese A-share listed companies in 2017, regardless of the time span from sample cover, or from the platform diversity, linguistic diversity point of view, there is a large room for improvement. Secondly, the optimization of the sentiment lexicon in this paper relies only on the different parameter combinations manually set, but how to find an optimal combination is still an unsolved problem. Finally, in the evaluation of stock market sentiment lexicon, only the single factor of portfolio return rate is considered, and other factors such as fluctuations in portfolio return are not taken into consideration, which is the direction that can be further studied in the future.

## References

1. Antweiler, W., Frank, M.Z.: Is all that talk just noise? The information content of internet stock message boards. *J. Finan.* **59**(3), 1259–1294 (2004). <https://doi.org/10.1111/j.1540-6261.2004.00662.x>

2. Bollen, J., et al.: Twitter mood predicts the stock market. *J. Comput. Sci.* **2**(1), 1–8 (2011). <https://doi.org/10.1016/j.jocs.2010.12.007>
3. Challa, M.L., et al.: Forecasting risk using auto regressive integrated moving average approach: an evidence from S&P BSE Sensex. *Finan. Innov.* **4**(1), 24 (2018). <https://doi.org/10.1186/S40854-018-0107-Z>
4. Koppel, M., Shtrimerberg, I.: Good news or bad news? Let the market decide. In: Shanahan, J. G., et al. (eds.) *Computing Attitude and Affect in Text: Theory and Applications*, pp. 297–301. Springer, Dordrecht (2006). [https://doi.org/10.1007/1-4020-4102-0\\_22](https://doi.org/10.1007/1-4020-4102-0_22)
5. Li, Q., et al.: Media-aware quantitative trading based on public Web information. *Decis. Support Syst.* **61**, 93–105 (2014). <https://doi.org/10.1016/j.dss.2014.01.013>
6. Li, Q., et al.: The effect of news and public mood on stock movements. *Inf. Sci.* **278**, 826–840 (2014). <https://doi.org/10.1016/j.ins.2014.03.096>
7. Liu, B.: Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **5**(1), 1–167 (2012). <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>
8. Liu, Y., et al.: A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. *Inf. Sci.* **394–395**, 38–52 (2017). <https://doi.org/10.1016/j.ins.2017.02.016>
9. Liu, Y., et al.: A method for ranking products through online reviews based on sentiment classification and interval-valued intuitionistic fuzzy TOPSIS. *Int. J. Inf. Tech. Decis. Making* **16**(6), 1497–1522 (2017). <https://doi.org/10.1142/S021962201750033X>
10. Loughran, T., Mcdonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finan.* **66**(1), 35–65 (2011). <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
11. Mahendhiran, P.D., Kannimuthu, S.: Deep learning techniques for polarity classification in multimodal sentiment analysis. *Int. J. Inf. Tech. Decis. Making* **17**(3), 883–910 (2018). <https://doi.org/10.1142/S0219622018500128>
12. Rao, H., et al.: Automatic construction of financial semantic orientation lexicon from large-scale Chinese news corpus. *Institut Louis Bachelier* **20**(2), 1–18 (2014)
13. Nayak, S.C., Misra, B.B.: Estimating stock closing indices using a GA-weighted condensed polynomial neural network. *Finan. Innov.* **4**(1), 21 (2018). <https://doi.org/10.1016/j.dss.2016.02.013>
14. Oliveira, N., et al.: Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decis. Support Syst.* **85**, 62–73 (2016). <https://doi.org/10.1186/S40854-018-0104-2>
15. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Comput. Linguist.* **35**(2), 311–312 (2009). <https://doi.org/10.1162/coli.2009.35.2.311>
16. Rashid, A., Jabeen, N.: Financial frictions and the cash flow – external financing sensitivity: evidence from a panel of Pakistani firms. *Finan. Innov.* **4**(1), 15 (2018). <https://doi.org/10.1186/S40854-018-0100-6>
17. Rosenthal, S., et al.: SemEval-2014 task 9: sentiment analysis in Twitter. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 73–80. Association for Computational Linguistics (2015). <https://doi.org/10.3115/V1/S14-2009>
18. Schumaker, R.P., et al.: Evaluating sentiment in financial news articles. *Decis. Support Syst.* **53**(3), 458–464 (2012). <https://doi.org/10.1016/j.dss.2012.03.001>
19. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: the AZFin text system. *ACM Trans. Inf. Syst.* **27**, 29 (2009)
20. Shleifer, A., Summers, L.H.: The noise trader approach to finance. *J. Econ. Perspect.* **4**(2), 19–33 (1990). <https://doi.org/10.1257/jep.4.2.19>
21. da Silva, N.F.F., et al.: Tweet sentiment analysis with classifier ensembles. *Decis. Support Syst.* **66**, 170–179 (2014). <https://doi.org/10.1016/j.dss.2014.07.003>

22. Song, Y., et al.: Sustainable strategy for corporate governance based on the sentiment analysis of financial reports with CSR. *Finan. Innov.* **4**(1), 2 (2018). <https://doi.org/10.1186/S40854-018-0086-0>
23. Sun, Y., et al.: A novel stock recommendation system using Guba sentiment analysis. *Pers. Ubiquit. Comput.* **22**(3), 575–587 (2018). <https://doi.org/10.1007/s00779-018-1121-x>
24. Turney, P.D., Littman, M.L.: Measuring praise and criticism: inference of semantic orientation from association. *ACM Trans. Inf. Syst.* **21**(4), 315–346 (2003). <https://doi.org/10.1145/944012.944013>
25. Wang, N., et al.: Textual sentiment of Chinese microblog toward the stock market. *Int. J. Inf. Technol. Decis. Making (IJITDM)* **18**(02), 649–671 (2019). <https://doi.org/10.1142/S0219622019500068>
26. Yousaf, I., et al.: Herding behavior in Ramadan and financial crises: the case of the Pakistani stock market. *Finan. Innov.* **4**(1), 16 (2018). <https://doi.org/10.1186/S40854-018-0098-9>
27. Yuen, R.W.M., et al.: Morpheme-based derivation of bipolar semantic orientation of Chinese words. In: *Proceedings of the 20th International Conference on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg (2004). <https://doi.org/10.3115/1220355.1220500>