Front. Eng. Manag. 2017, 4(2): 201–211
DOI 10.15302/J-FEM-2017018

RESEARCH ARTICLE

Xianjiao WU, Xiaolin WANG, Shudong MA, Qiang YE

The influence of social media on stock volatility

Abstract This study explores the influence of social media on stock volatility and builds a feature model with an intelligence algorithm using social media data from Xueqiu.com in China, Sina Finance and Economics, Sina Microblog, and Oriental Fortune. We find that the effect of social factors, such as increased attention to a stock's volatility, is more significant than public sentiment. A prediction model is introduced based on social factors and public sentiment to predict stock volatility. Our findings indicate that the influence of social media data on the next day's volatility is more significant but declines over time.

Keywords stock volatility, social data, sentiment analysis, boosting algorithm

1 Introduction

With the rapid development of social networks and the emergence of Internet finance, investors are paying more attention to social media platforms. Financial social platforms have become an effective channel for investors to interact with others and gain insights into financial market trends. The massive amounts of social media data available today can confer huge commercial and academic value.

Schumaker and Chen (2009a, 2009b) analyzed the financial news and put forward a prediction model based on machine learning. They investigated more than 9200 financial articles and over 10 million comments about S&P 500 stocks. Their findings demonstrated that a prediction model can estimate the stock price about 20 minutes later. By looking into the relationship between data from Raging Bull Bulletin Board System and stock returns, Tumarkin and Whitelaw (2001) demonstrated that there was no significant relationship between investor sentiment and stock returns. On the other hand, Tetlock (2007) and Tetlock et al. (2008) studied the "Abreast of the Market" column in the Wall Street Journal and indicated that negative factors in the press may lead to falling stock prices. Choi et al. (2000) conducted an experiment, and found that the volume of some companies trading on Internet doubled the volume of those companies trading without Internet. Freedman and Jin (2011) evaluated the credit information asymmetry with the information of lenders and borrowers on Prosper.com. Their findings contributed to reduce the information asymmetry and thereby reduced the risk of asset pricing. He et al. (2016) explored whether sentiments in "tweets" relate to stock price changes at these companies. Their findings indicated that negative sentiments predicted a firm's future stock prices. Such results prove the possible predictive value of social media data on stock prices at the company level. Oliveira et al. (2017) assessed the value of microblogs in forecasting stock market returns, volatility, and trading volume of diverse indices and portfolios. Their findings suggested that Twitter sentiment and posting volume are relevant for forecasting returns in the S&P 500 index, portfolios of lower market capitalization, and some industries. Tirunillai and Tellis (2012) employed multivariate time-series models to assess the relationship between user generated content and stock market performance. Their findings suggested that the volume of chatter significantly led abnormal returns by a few days. The effect of negative and positive metrics of UGC on abnormal returns was asymmetric.

With the advances in data mining and machine learning algorithms, researchers can now make even more discoveries in this field. Antweiler and Frank (2004) used semantic analysis to investigate the data from Raging Bull and Yahoo Finance BBS. They found that the Dow Jones average constituents were embedded in 45 samples and demonstrated the correlation between stock returns and the information index of BBS. Choudhury et al. (2008) tried to associate the stock market behavior with blog communication information. They employed a support vector

Received March 24, 2017; accepted June 1, 2017

Xianjiao WU, Xiaolin WANG (✉), Shudong MA, Qiang YE
School of Management, Harbin Institute of Technology, Harbin 150001, China
E-mail: wxl-jun@163.com

machine regression model with communication information from Engadget and successfully predicted stock price movement trends. Moreover, the experimental results showed that the prediction error rate was as low as 13% to 22%. Based on the analysis of the relationship between the extent of the news on public companies and their stock returns, Fang and Peress (2008) drew the conclusion that premium issue was more likely to happen to stocks that garnered little attention. Asur and Huberman (2010) successfully predicted the box-office sales through analyzing the content of the Twitter. Their findings demonstrated that the model constructed with the content and the number of content generated by user can forecast some things in life. In addition, Pak and Paroubek (2010) made contribution to the study of sentiment analysis and information dissemination.

In the context of finance and technology, Bollen et al. (2016) from Indiana University used two different mood-tracking tools to conduct a study of Twitter. OpinionFinder can categorize articles into positive and negative ones, while Google profile of mood states (GPOMS) is a new emotional testing tool based on the "Profile of Mood States" in clinical applications. GPOMS can help to divide user sentiments into six clusters, friendly, watchful, confident, happy, energetic, and calm. The results revealed that the index of "calm" was very close to the Dow Jones industrial average, and the Dow Jones industrial average lagged three days behind the index of "calm". In addition, the test of the self organizing fuzzy neuron network (SOFNN) prediction model found that the model without emotional factors had a 73.3% accuracy rating, while the model with emotional factors had an 86.7% accuracy rating.

Although the Internet contains extensive open-source information that can drive investor behavior, the difficulties in obtaining and processing data and the reliability of the data block studies in this field (Das and Chen, 2007). Most previous studies devoted to establishing the relationship between public sentiment and stock market movement use data from the Sina microblogs. However, there are many other factors that can affect the stock market. Additionally, microblogs do not embody a large number of emotional investor trends. This study explores which type of social media data has more significant influence on stock volatility. We introduce the prediction model with social media data and try to predict stock volatility for some time to come. Our data set includes a combination of stock rankings, stock portfolios, and public sentiment from the Xueqiu website, and public sentiment from microblogs, BBS, and other blogs.

The rest of the paper is organized as follows. We first introduce the context of our study and put forward our hypotheses. Section 3 describes our data set and method used for data processing. To find variables that have more significant effects on the volatility of stocks, a random forests model is introduced. Finally, we establish a prediction model with a boosting algorithm and examine the correlation between the social data and stock volatility.

2 Background and hypothesis

An extensive amount of literature explores the factors of stock volatility. In this study, social media data include both communication and interaction information on the Internet such as investor opinions of the stock market or their own investments and transactions. Our data consist of messages and comments from Sina Microblog, other blogs, and BBS, with attention to specific stocks or the market and information on stock portfolios. In such cases, investors are able to scan other investor opinions and transactions before making investment decisions. Accordingly, they are susceptible to such sentiments. On the other hand, investor sentiments can dominate investment intention, and thereby affect the stock market. Based on a literature review, we find that most previous studies use text mining and sentiment analysis of microblogs. Few studies focus on other open-source Internet information. Therefore, public sentiment information in Sina Microblog may be more likely to significantly affect stock market movement than other open-source Internet information. We thus propose the following:

Hypothesis 1. The influence of public opinion information on stock volatility is more significant than that of open-source Internet information (such as new attention on stocks, stock portfolio information, etc.)

The emotional tendency of investors in the stock market is an important index of investor sentiment. Cheng and Lin (2013) points out in her study that in a social network, the investor sentiment index positively correlates with both the stock index return and volume. Specifically, stock index returns and volume have a significant effect on the investor sentiment index of 40 trading days, while the investor sentiment index influences the stock index return in reverse for a short period. Pang et al. (2012) and Cheng and Lin (2013) each carried out a study on the Sina Microblog, and both studies proved that sentiment embedded in a microblog had a certain influence on the fluctuation in the stock market. Wen's research (Wen et al., 2014) revealed that investors' positive attitudes had an influence on stock returns. However, the influence of negative attitudes was not significant. Wen explained that sense played a leading role in investment when investors held negative attitudes towards the market. The results indicate that personal sentiment could be influenced by other investors and the stock market, and has a significant impact on asset allocation strategy. Therefore, public opinion in web text can be used as an index of stock price fluctuations. Liu et al. (2011) used a search volume index (SVI) to predict stock volatility, and her finding supports that a SVI can effectively forecast the annual yield of the Shanghai Composite Index. By looking into the relationship between the growth in Internet searching and relative stock prices, Yang et al. (2013) concluded that the spread of information on the Internet could affect stock price fluctuations. With the help of data mining, Feng (2013) performed research on the influence of open-source Internet information on stock volatility. Relying on a macroeconomic data set and the Baidu Index, she confirms the strong correlation between Internet information and stock market price fluctuations. Gao (2009) revealed the infectivity of sentiment in social networks in her study. Analogously, we assume that infectivity exists on financial social platforms.

The data involved in this study are open-source Internet information, including attention rankings, transaction rankings, and information about stock portfolios. Based on the preliminary analysis and conclusions, the spread of these data on social networks probably induce stock price fluctuation. We thus propose the following:

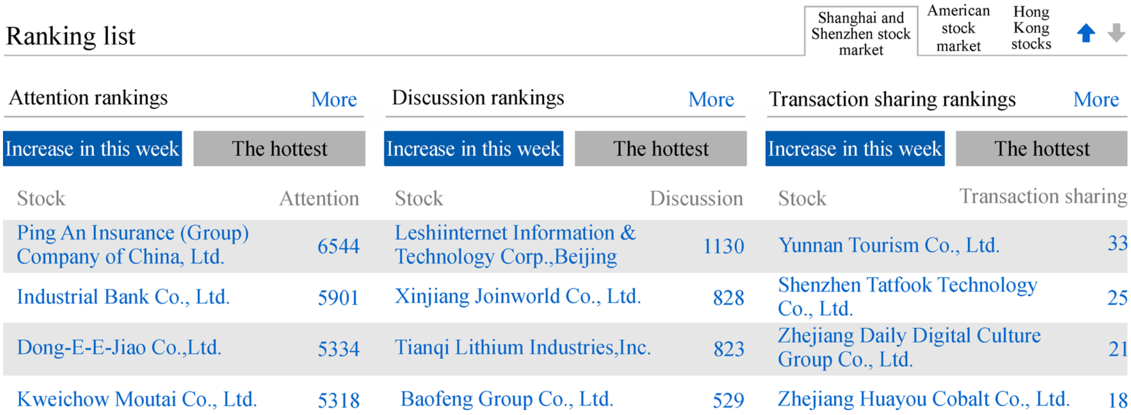
Hypothesis 2. A combination of public sentiment and social data (such as stock rankings and stock portfolio information) can predict stock prices to a certain precision.

Hypothesis 3. The impact of social data on stock prices can linger for some time.

3 Methodology

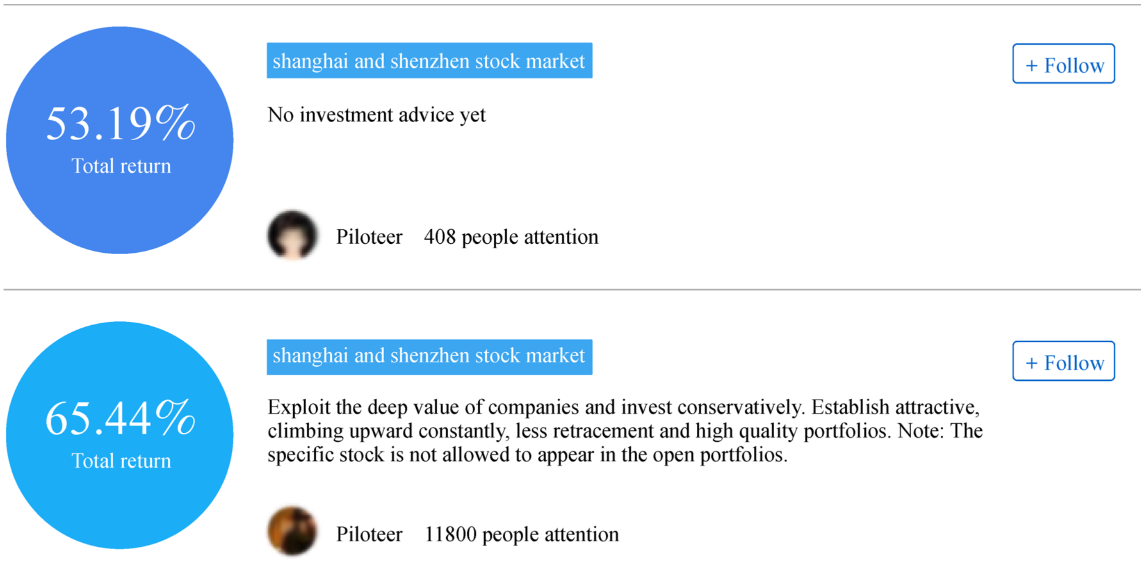
3.1 Data collection

Xueqiu.com is a professional financial social platform with the highest participation in China. It is a perfect data source that can provide us with very valuable information including rankings and stock portfolio data (Figs. 1 and 2). We also obtain basic stock information from China Economic Net, including stock ticker, stock name, stock value, and industry. In addition, a crawler is used to gather data regarding public opinion on the sites Xueqiu.com, Sina Finance and Economics, Sina Microblog, and



Stock	Attention	Stock	Discussion	Stock	Transaction sharing
Ping An Insurance (Group) Company of China, Ltd.	6544	Leshiinternet Information & Technology Corp.,Beijing	1130	Yunnan Tourism Co., Ltd.	33
Industrial Bank Co., Ltd.	5901	Xinjiang Joinworld Co., Ltd.	828	Shenzhen Tatfook Technology Co., Ltd.	25
Dong-E-E-Jiao Co.,Ltd.	5334	Tianqi Lithium Industries,Inc.	823	Zhejiang Daily Digital Culture Group Co., Ltd.	21
Kweichow Moutai Co., Ltd.	5318	Baofeng Group Co., Ltd.	529	Zhejiang Huayou Cobalt Co., Ltd.	18

Fig. 1 Translated screen shot of Xueqiu.com — rankings data



53.19%
Total return
shanghai and shenzhen stock market
+ Follow

No investment advice yet

Piloteer 408 people attention

65.44%
Total return
shanghai and shenzhen stock market
+ Follow

Exploit the deep value of companies and invest conservatively. Establish attractive, climbing upward constantly, less retracement and high quality portfolios. Note: The specific stock is not allowed to appear in the open portfolios.

Piloteer 11800 people attention

Fig. 2 Translated screen shot of Xueqiu.com — stock portfolio data

Oriental Fortune. These data represent factors influencing stock volatility. However, the stock price data are essential for training data in a supervised way. Hence, we acquire the stock price from Xueqiu.com (Fig. 3). The crawler is used for data collection, which is written in Python. Before entering the data in the MySQL database, we have to remove noise and handle missing values for the data cleaning.

3.2 Structured handling of non-text data

In order to look into the relationship between each stock's volatility and the social data, we need to show detailed data for each stock. Thus, data refactoring and data preprocessing are essential procedures.

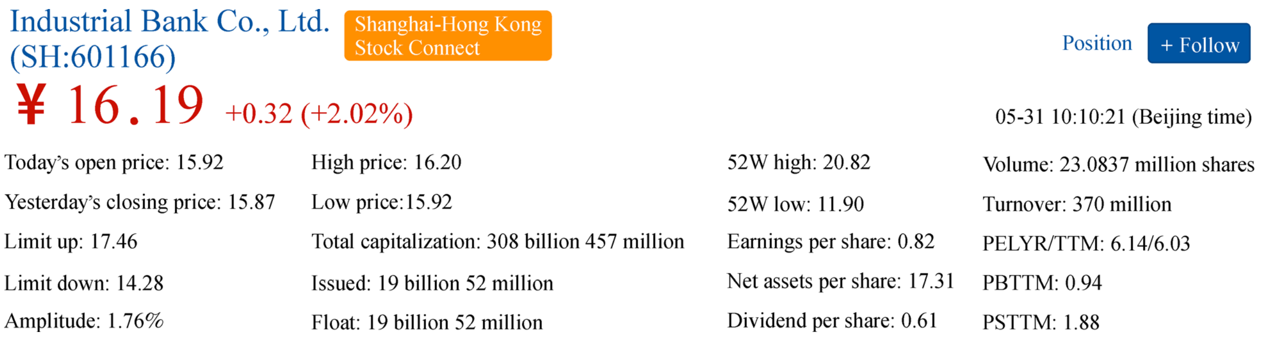
First, this research adopts the min-max method to standardize the sharing ranking. Standardizing reduces the errors due to different dimensions of variables and makes the data available for feature selection and model building. Table 1 presents the standardized data attention sample.

Second, we reconstruct the stock portfolio data to show information on each stock. Table 2 shows the stock portfolio data after refactoring.

Finally, we need to classify the stock price trend as the stock serves as the dependent variable in the prediction model. In this study, we extract data on a certain day (t), and choose the stock price on $t+1$, $t+2$, and $t+3$ as dependent variables to explore how long the influence of the social data on stock volatility lasts.

We thus classify the stock price trend in two ways. First, the stock price trend is divided into "advancing" and "declining" as in Eq. 1:

$$\text{The trend of the stock price} = \begin{cases} 1(\text{volatility} \geqslant 0) \\ -1(\text{volatility} < 0) \end{cases}. \tag{1}$$



Industrial Bank Co., Ltd. (SH:601166) Shanghai-Hong Kong Stock Connect

¥ 16.19 +0.32 (+2.02%)

05-31 10:10:21 (Beijing time)

Today's open price: 15.92	High price: 16.20	52W high: 20.82	Volume: 23.0837 million shares
Yesterday's closing price: 15.87	Low price:15.92	52W low: 11.90	Turnover: 370 million
Limit up: 17.46	Total capitalization: 308 billion 457 million	Earnings per share: 0.82	PELYR/TTM: 6.14/6.03
Limit down: 14.28	Issued: 19 billion 52 million	Net assets per share: 17.31	PBTTM: 0.94
Amplitude: 1.76%	Float: 19 billion 52 million	Dividend per share: 0.61	PSTTM: 1.88

Fig. 3 Translated screen shot of Xueqiu.com — Real-time basic stock data

Table 1 Summary statistics of attention by company

Stock_id	SH600886	SZ002702	SZ002276	SZ000025	SZ002751
Company references	12516	6741	6256	5911	5593
Company references(standardized)	1	0.54	0.5	0.47	0.45

Table 2 Stock portfolio variable descriptions after refactoring

Variables	Description
Stock ticker	Unique identification of stock
Share of stock in portfolios	Share of stock in top 100 portfolios of a category
Daily return of stock in portfolios	Daily return of stock shown in top 100 portfolios
Monthly return of stock in portfolios	Monthly return of stock shown in top 100 portfolios
Annual return of stock in portfolios	Annual return of stock shown in top 100 portfolios
Total yield in portfolios	Total yield of stock shown in top 100 portfolios
Investment style of portfolios	Investment style of stock shown in top 100 portfolios
Attention of stock in portfolios	Attention of stock in portfolios
Date	Date when occurred

Second, the stock price trend is divided into “advancing,” “flat,” and “declining” as in Eq. 2:

$$\text{The trend of the stock price} = \begin{cases} 1(\text{volatility} \geqslant 0.002) \\ 0(-0.002 < \text{volatility} < 0.002) \\ -1(\text{volatility} \leqslant -0.002) \end{cases}. \tag{2}$$

3.3 Sentiment analysis of text data

In this study, we employ a dictionary-based approach for sentiment analysis on public opinion. The process of sentiment analysis involves the establishment of a corpus, the extraction of feature words, the establishment of an emotional dictionary, and the assessment of emotional tendencies.

A corpus is needed before sentiment analysis. The first step is word segmentation with the help of the jieba plug-in written in Python. The Chinese corpus employed in this research is dict.txt.big from jieba, while the stopword list is from oschina. However, there are many specialized vocabularies on the stock market that cannot be segmented by the dictionary and must be replenished artificially. Second, we need to specify sentiment words that represent the stock fluctuation. We then filter out the emotional words from the segmented words. The final step involves assessing these words artificially, categorizing them into “advancing,” or “declining” and adding these words to the defined corpus.

The next phase is to extract feature words. After each round of word segmentation, the term frequency–inverse document frequency (TF/IDF) algorithm is utilized to extract the top K keywords that can represent the sentiment fragment, and write keywords into the database to build the emotional dictionary (Salton and Mcgill, 1983). On the basis of the article keywords, the emotional words are filtered out. We prefer a “bag” of words to a set of words model as the former counts term frequency (Sivic and Zisserman, 2009). A model incorporating term frequency is more accurate.

Finally, we assess the emotional polarity of an essay based on the emotional dictionary after emotional word extraction. The essays are categorized into “positive,” “negative,” and “neutral” according to Eq. 3 and the categorized results are shown in Table 3.

$$\text{Emotional polarity of an essay} = \begin{cases} \text{positive, if the number of positive words} \\ \qquad > \text{the number of negative words} \\ \text{neutral, if the number of positive words} \\ \qquad = \text{the number of negative words} \\ \text{negative, if the number of positive words} \\ \qquad < \text{the number of negative words} \end{cases} \tag{3}$$

To determine the sentiment tendency towards the stock market, emotional polarity is quantified as “1” for “positive,” “0” for “neutral,” and “– 1” for “negative.” All the articles are summarized on the same day, and the sentiment tendency towards the stock market is determined as in Eq. 4:

$$\begin{aligned} &\text{The emotional tendency of stock market} \\ &= \sum(\text{the emotional tendency of each article} \\ &\quad \div \text{ the number of articles}) > 0.5?1 : \qquad . \\ &\quad \sum(\text{the emotional tendency of each article} \\ &\quad \div \text{ the number of articles}) < -0.5?-1 : 0 \end{aligned} \tag{4}$$

The quality of the text analysis is mainly reflected in three aspects: the first is the result of word segmentation, the second is whether emotional words are completely extracted, and the third is the quality of the sentiment analysis algorithm. Therefore, we randomly select 1000 articles, and compare the emotional polarity of articles with the predicting results. Figure 4 shows the accuracy curve of sentiment classification as the number of articles increases.

In Fig. 4, the accuracy of sentiment classification eventually grows to about 0.9, which indicates that the classification model is basically successful.

Table 3 An example of emotional polarity assessment

Attribute	Description
Segment of an essay	“In the first week after the Spring Festival, the A-shares began to rebound as expected, and some specific stocks had excessive growth within a few days. In the wake of the rebound, the prices of A-shares fluctuate slightly on Friday. The Shanghai Composite Index closed at 2860.02 with a decline of 0.10%, while the Growth Enterprise Index rose by 0.92%. The stocks of shipbuilding, electronic information industry, the water and gas supplier, the electronic components, the foreign trade industry, the glass industry, the food industry, and the aircraft industry rose more than stocks of petroleum industry, the real estate, the steel industry, and the brewing industry.”
Emotional word extraction	Rebound, growth, rebound, fluctuate, decline, rose
Emotional polarity	Positive

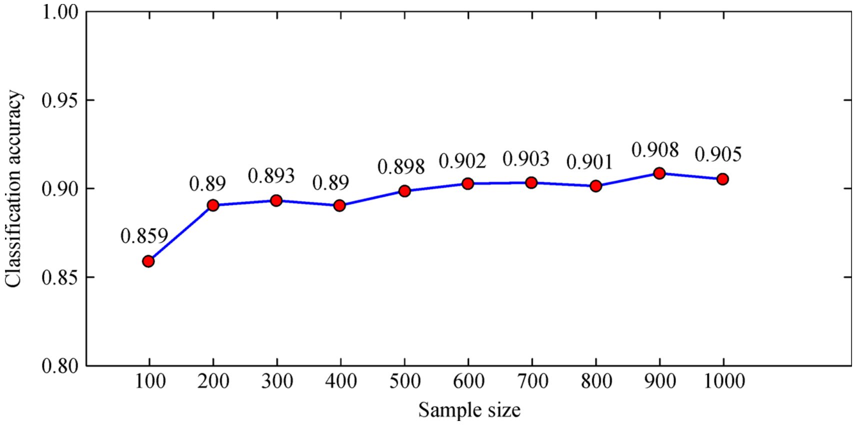


Fig. 4 Sketch of sentiment classification training effect

4 Results

4.1 Influence factors of stock volatility

In empirical research, there are always excessive characteristic variables, some of which may be redundant. Feature selection is essential to reduce spare variables, improve the effect of classification, and enhance the explanatory power of characteristic variables. Therefore, the correlational analysis between social data (such as new attention, discussion, transaction sharing data) and volume is conducted. The results of the correlational analysis and significance test are presented in Table 4.

The results in Table 4 support that there is a high correlation between stock attention, discussion, the amount of transaction sharing, and the stock volume. According to previous literature, the volume has a significant impact on stock volatility (Li et al., 2007; Shi, 2005). Thus, we conclude that social data, like stock attention, influences stock volatility.

The analysis of public sentiment shows that the proportion of “positive” attitudes is associated with the fluctuation of the market. According to the emotional words extracted from articles between March 1, 2016 and March 31, 2016, there are more positive words than negative words. The frequency of the emotional word also reflects the rebound or rise in the stock price. This inference is confirmed in Fig. 5, where the Shanghai Composite Index is rising most of the time between March 1, 2016 and March 31, 2016. That is, the emotional polarity and proportion of emotional words embodied in social public opinion reflect the stock market fluctuation over a period of time.

The research on portfolio information proposed by investors suggests that the portfolio data has little direct correlation with stock volatility. However, portfolio information on social platforms receives extensive attention and provides investors with a reference. Considering that it may work together with other variables to influence the stock price, we add it as a feature variable.

In order to confirm the validity of Hypothesis 1, this study introduces extremely randomized trees to perform tests on the importance of the characteristic variables (Geurts et al., 2006). The criterion that extremely randomized trees employs to assess the importance of the feature variables is the Gini-impurity. Gini-impurity represents the possibility of a random sample being categorized into the fault subset, and the importance of the feature variable is measured by the increase in model Gini-impurity due to removal of the variable. Figure 6 depicts the importance of the features in this study (Yang et al., 2014).

As shown in Fig. 6, the most important feature variables refer to the basic characteristics of the stock such as

Table 4 The correlational analysis between social data and volume

Variables	Description	Correlation	Significance
Attention increases this week	Attention increases in the stock this week	0.619**	0.0021
Attention is hottest	Attention of the stock in the hottest list	0.418**	0.0049
Discussion increases this week	Discussion increases on the stock this week	0.406**	0.0052
Discussion is hottest	Discussion of the stock in the hottest list	0.469**	0.0040
Transaction sharing increases in this week	Transaction sharing increases on the stock this week	0.364**	0.0089
Transaction sharing is hottest	The amount of transaction sharing of the stock in the hottest list	0.547**	0.0037

Note: **denotes statistical significance at the 1% level; *denotes statistical significance at the 5% level

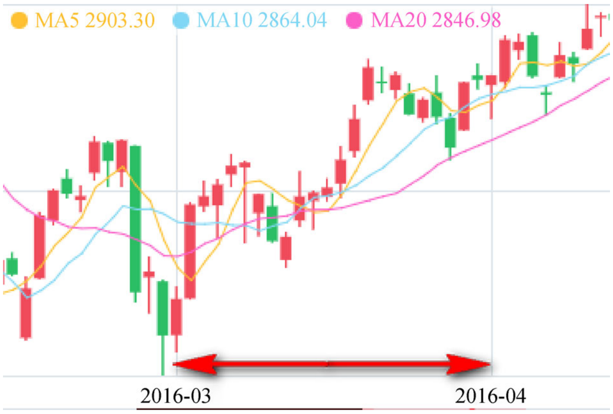


Fig. 5 Daily chart of the Shanghai Composite Index

volume and turnover rate. The social features are less important than basic characteristics, but still more important than variables about public opinion information and portfolios. Therefore, we can generally sort the feature variables in the order of their importance. As stated, the most important variables are the basic characteristics, followed by social data, some features of the portfolios, public opinion information, and the other features of the portfolios.

To rule out the influence of basic characteristics of the stocks and highlight the importance of the social data, we remove the basic stock characteristics and variables that are not significant. Figure 7 delineates the analysis results with only social data considered.

The results presented in Fig. 7 are nearly consistent with Fig. 6. The influence of "attention increase this week," "transaction sharing the hottest," "discussion increase this week," "attention the hottest," and "discussion the hottest" on stock volatility is highly significant. According to the results of the feature selection, some open-source data have

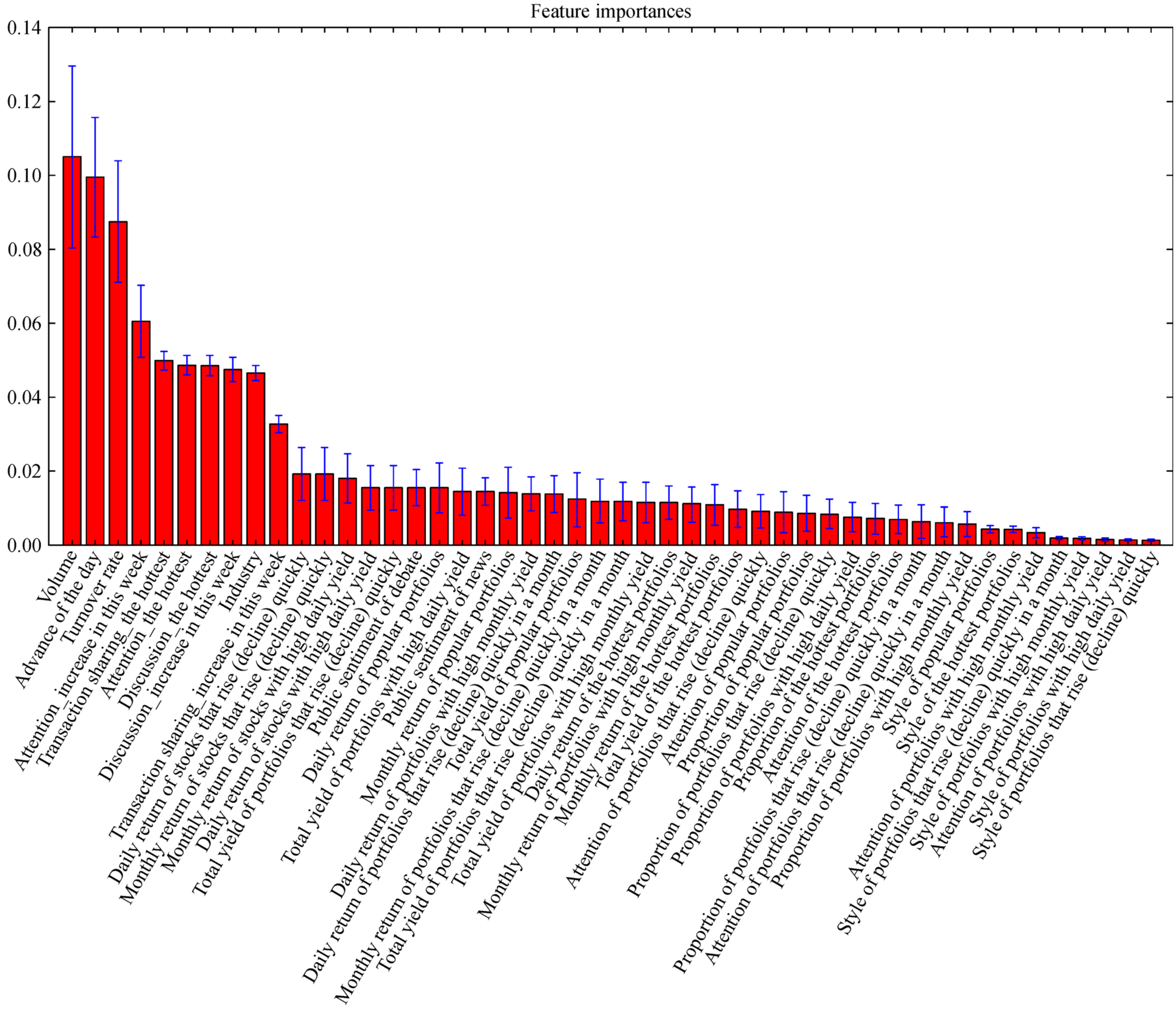


Fig. 6 The importance of features

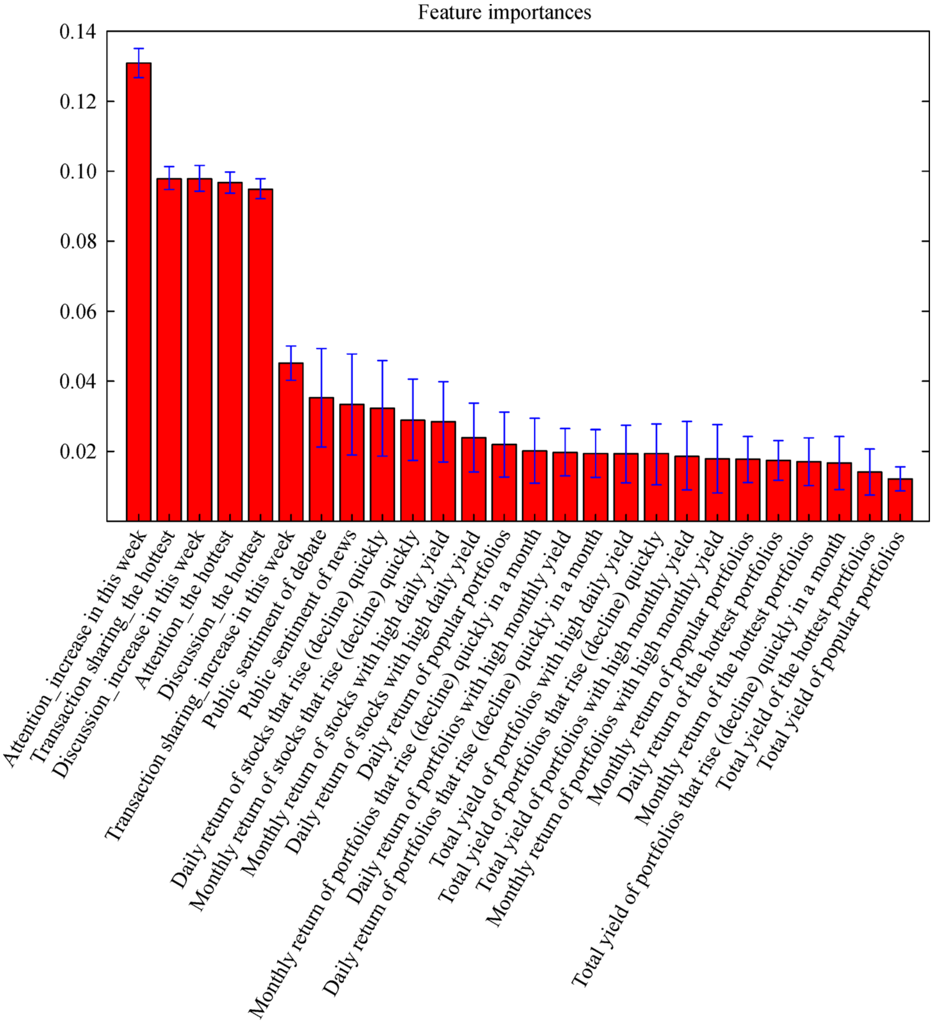


Fig. 7 The importance of social attributes

more influence on the stock price than public sentiment. Therefore, Hypothesis 1 is rejected. Most previous studies have ignored the value of social data.

Based on the results of feature selection, we add important feature variables to the prediction model. In order to improve the prediction accuracy and highlight the value of social data, we establish two prediction models to examine the correlation between social data and stock price. One model includes basic characteristics of the stocks (such as volume, turnover rate, etc.) while the other does not.

4.2 Prediction model of stock volatility

To explore whether the prediction model can predict the stock price trend successfully, different algorithms are employed to confirm the validity of Hypotheses 2 and 3 based on variable selection.

There are three kinds of factors that can predict the stock volatility. These are social data, public sentiment, and portfolio information. The movement of stock price serves as the dependent variable in the model. The tests of the different algorithms imply that classification performs better than regression. Therefore, the classification prediction model is established to assess stock volatility and the algorithm is adopted to train data. A conceptual model is shown in Fig. 8.

In this study, we collect data from October 1, 2015 to April 11, 2016. It is worth noting that the circuit breaker came into effect on January 1, 2016 in China, and came to an end on January 8, 2016. Therefore, the stock index fluctuated dramatically during the first month of 2016, and the stock index dropped sharply on February 25, 2016. We randomly select 80% data as training data. The final data set contains 186995 records after excluding missing values and data that have no direct impact on stock volatility.

To test and assess the model, we employ precision, recall rate (P/R), and F1 as assessment indicators and draw the

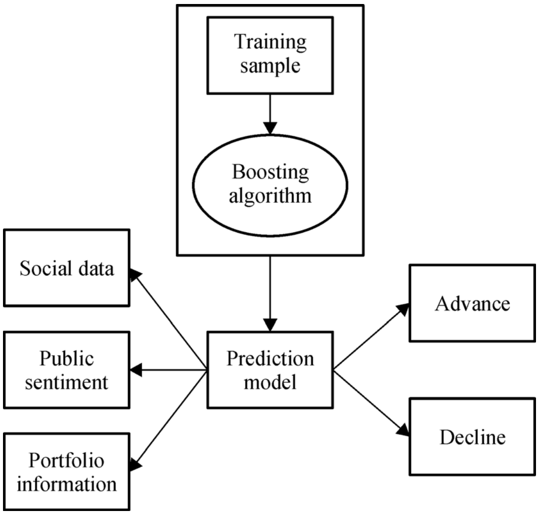


Fig. 8 Conceptual model

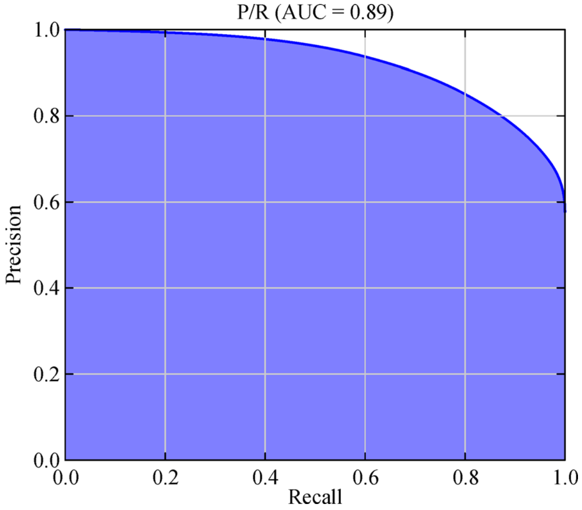


Fig. 9 Training social data P/R curve

P/R curve to calculate area under the curve (AUC) (Zou et al., 2007). Three assessment indicators are calculated as in Eqs. 5, 6, and 7:

$$\text{Precision} = \frac{|\cap(\text{PredictionSet},\ \text{ReferenceSet})|}{\text{PredictionSet}}, \tag{5}$$

$$\text{Recall} = \frac{|\cap(\text{PredictionSet},\ \text{ReferenceSet})|}{\text{ReferenceSet}}, \tag{6}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

where PredictionSet represents the stock set that is predicted to rise (decline), and ReferenceSet represents the stock set that rises (declines) in the prediction set.

The experimental results show that the accuracy of the gradient boosting algorithm is higher. Random sampling is used as the main classification method, and cross validation as the auxiliary. We remove the non-social data to highlight the influence of the social data, and then conduct training and testing on the data. The assessment indicators are shown in Table 5 and Fig. 9.

The average accuracy of the training in practice is 0.805, and the value of the AUC turns out to be 0.89, which indicates that social data can predict stock volatility.

On a sample of stock prices from February 18, 2016 to March 8, 2016, we conduct the prediction model in a real environment. It should be noted that data in the period of the prediction should be removed from the training sample. The predicting effect of the model based on the gradient boosting algorithm is shown in Table 6:

Table 5 The precision, recall rate, F1, and support of social data

Class	Precision	Recall	F1-score	Support
Decline	0.77	0.77	0.77	22983
Rise	0.83	0.83	0.83	30644
Avg/total	0.81	0.81	0.81	53627

Table 6 shows that predicting accuracy is generally over 0.6 except on February 25, 2016, and is higher on certain days, reaching even 0.8 or 0.9. The model can forecast the stock price trend to some extent. Furthermore, the accuracy on February 25, 2016 is 0.164 as the Shanghai Composite Index plunged 6.41% on that day. Thus, the prediction model cannot predict abnormalities well.

The accuracy and real effect of the prediction model prove that social data can predict a stock price trend to some extent. Thus, Hypothesis 2 is supported.

We have proved the influence of social data on stock volatility from the analysis of the stock price in $t + 1$. The "volume" is added into the model to improve the accuracy of the prediction in the following experiment. To look into the duration of the impact of social data on stock volatility and test Hypothesis 3, we model the stock price in $t + 1$, $t + 2$, and $t + 3$ as the dependent variable. The accuracy of the models is shown in Table 7.

As can be seen, the influence of social data on stock volatility can last several days, but the strength of the influence declines with time. Thus, Hypothesis 3 is supported.

5 Conclusions

Unlike previous studies on web information and the stock market, this study focuses on the analysis of social data rather than public opinion. We make a comparison of the importance of various social attributes for the stock market. The empirical results suggest that public sentiment is consistent with the stock market trend over a period of time. The social data, such as attention, discussion, and

Table 6 Predicting accuracy of stock volatility in $t+1$

Predicting date	2.18	2.19	2.22	2.23	2.24	2.25	2.26
Accuracy	0.841	0.649	0.943	0.613	0.586	0.164	0.675
Predicting date	2.29	3.1	3.2	3.3	3.4	3.7	3.8
Accuracy	0.808	0.805	0.839	0.577	0.678	0.849	0.621

Table 7 The predicting effect in $t+1$, $t+2$, and $t+3$

Date	$t+1$	$t+2$	$t+3$
Cross validation accuracy	0.67	0.52	0.47
Random sampling accuracy	0.86	0.85	0.82

sharing, can reflect the emotion of investors. The result of feature selection indicates that the influence of social data on stock price is more significant than public sentiment. Notably, the information on specific stocks in portfolios does not have a direct impact on stock volatility. However, it can improve the accuracy of the prediction model based on the boosting algorithm. This research finally proves that a prediction model using social data based on random sampling can predict stock volatility. Moreover, the influence can last three days, but then declines gradually. However, the model does not work when emergencies or abnormalities occur.

Social financial platforms on the Internet make it easier and more convenient for users to perceive other investors' attitudes to the stock market more precisely. Furthermore, the infectivity of social sentiment affects other users of the same social platform. The study on the correlation between social data and stock volatility can contribute to the development of the platform as well as to the decisions of investors. This study contributes to the knowledge of correlation between social data and stock volatility. In previous studies, the main focus has been on emotions in microblogs. Adding open-source social data of the Internet to the prediction model here provides a new method for future studies. Furthermore, this study enriches the research in this field by testing the influence of social data on stock volatility. In practice, this study proposes a method to predict stock volatility, which can help investors allocate assets appropriately.

This study does have some limitations. First, it collects data from September 2015 to April 2016, which hardly covers all the attributes of stock volatility. Second, data are mainly collected from Xueqiu.com. Although Xueqiu.com has become the dominant social platform, its unstable server leads to missing data. Finally, social attributes involved in this study are limited, and more social factors can be added to the model in the future.

Acknowledgements The author appreciates the comprehensive assistance of Mr. Shudong MA with data gathering and compilation that led to the drafting of this article. This research is supported by National Natural Science Foundation of China (Grant No. 71532004).

References

Antweiler W, Frank M Z (2004). Is all that talk just noise? The information content of Internet stock message boards. Journal of Finance, 59(3): 1259–1294

Asur S, Huberman B A (2010). Predicting the future with social media. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). IEEE, 1: 492–499

Bollen J, Mao H, Zeng X (2016). Twitter mood predicts the stock market. Eprint Arxiv, 2(1): 1–8

Chen Y (2016). Predicting stock trading volume through social media data. 2016-04-01, https://scholarworks.bridgeport.edu/xmlui/handle/123456789/1649

Cheng W Y, Lin J (2013). The relationship between stock index and investor sentiment in social media. Management Science, 26(5): 111–119

Choi J J, Laibson D, Metrick A (2000). Does the Internet increase trading? Evidence from investor behavior in 401(k) plans. Ssrn Electronic Journal, 64(12): 10–11

Choudhury M D, Sundaram H, John A, Seligmann D D (2008). Can blog communication dynamics be correlated with stock market activity? Hypertext 2008, Proceedings of the ACM Conference on Hypertext and Hypermedia, Pittsburgh, PA, USA, 55–60

Das S R, Chen M Y (2007). Yahoo! for Amazon: sentiment extraction from small talk on the web. Management Science, 53(9): 1375–1388

Fang L, Peress J (2008). Media coverage and the cross-section of stock returns. Social Science Electronic Publishing, 64(5): 2023–2052

Feng L N (2013). A Study on Influence of Open Source Information Flow on Stock Volatility. Dissertation for Master's Degree. Tianjin: Tianjin University

Freedman S, Jin G Z (2011). Learning by doing with asymmetric information: evidence from prosper. com. Nber Working Papers, 2011: 203–212

Gao X P (2009). The infectivity of emotion in social network. Pictorial of Science, (9): 20–22

Geurts P, Ernst D, Wehenkel L (2006). Extremely randomized trees. Machine Learning, 63(1): 3–42

He W, Guo L, Shen J, Akula V (2016). Social media-based forecasting: a case study of tweets and stock prices in the financial services industry. Journal of Organizational and End User Computing, 28(2): 74–91

Li Y L, Li S C, Yang G H (2007). The correlation between stock market volume and price. Journal of Hebei University of Economy and Trade, 28(2): 65–70

Liu Y, Lv B F, Peng G (2011). Predictive power of internet search data for stock market: a theoretical analysis and empirical test. Economic

Management, 33(1): 172–180

Oliveira N, Cortez P, Areal N (2017). The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications, 73: 125–144

Pak A, Paroubek P (2010). Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of the International Conference on Language Resources and Evaluation. Valletta, Malta

Pang L, Li S S, Zhang H (2012). Research on stock investor sentiment tendency based on microblog. Computer Science, B06: 249–252

Salton G, Mcgill M J (1983). Introduction to modern information retrieval. Library Management, 32(4/5): 373–374

Schumaker R P, Chen H (2009a). Textual analysis of stock market prediction using breaking financial news: the AZFin text system. ACM Transactions on Information Systems, 27(2): 1–19

Schumaker R P, Chen H (2009b). A quantitative stock prediction system based on financial news. Information Processing & Management, 45 (5): 571–583

Shi M J (2005). An analysis of volume's impact on stock yield. Statistics & Information Forum, 20(2): 60–62

Sivic J, Zisserman A (2009). Efficient visual search of videos cast as text retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(4): 591–606

Tetlock P C (2007). Giving content to investor sentiment: the role of media in the stock market. Journal of Finance, 62(3): 1139–1168

Tetlock P C, Saar-Tsechansky M, Macskassy S (2008). More than words: quantifying language to measure firms' fundamentals. Journal of Finance, 63(3): 1437–1467

Tirunillai S, Tellis G J (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. Marketing Science, 31(2): 198–215

Tumarkin R, Whitelaw R F (2001). News or noise? Internet postings and stock prices. Prehospital and Disaster Medicine, 57(3): 41–51

Wen F H, Xiao J L, Huang C X (2014). Research on influence of investors sentiment on stock price. Journal of Management Science, 17(3): 60–69

Yang D J, Yang J, Zhan X J (2014). Feature selection algorithm based on the random forest. Journal of Jilin University: Engineering Science, 44(1): 137–141

Yang X, Lv B F, Peng G (2013). The influence of emergency on stock market: an analysis based on online searching. Practice and Knowledge of Mathematics, 43(23): 17–28

Zou K H, O'Malley A J, Mauri L (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation, 115(5): 654–657