frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 30 August 2022
DOI 10.3389/fphy.2022.987799

OPEN ACCESS

EDITED BY
Wei Wang,
Chongqing Medical University, China

REVIEWED BY
Genhua Hu,
Anhui University of Technology, China
Mingjie Wang,
University of Guelph, Canada

*CORRESPONDENCE
Jinghua Tan,
jinghuatan.swufe@gmail.com

SPECIALTY SECTION
This article was submitted to Social Physics,
a section of the journal
Frontiers in Physics

RECEIVED 06 July 2022
ACCEPTED 28 July 2022
PUBLISHED 30 August 2022

CITATION
Zhang H, Chen Y, Rong W, Wang J and Tan J (2022), Effect of social media rumors on stock market volatility: A case of data mining in China.
Front. Phys. 10:987799.
doi: 10.3389/fphy.2022.987799

Effect of social media rumors on stock market volatility: A case of data mining in China

Hua Zhang[1], Yuanzhu Chen[2], Wei Rong[3], Jun Wang[4] and Jinghua Tan[4]*

[1]School of Economics and Management, Sichuan Normal University, Chengdu, China, [2]School of Computing, Queen's University, Kingston, ON, Canada, [3]School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu, China, [4]School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, China

The Stock Market is a typical complex network composed of investors, stocks, and market information. The abnormal fluctuation of the Stock Market has a strong effect on the economy of a country and even that of the world. Fueled by the herd effect of the increasingly abundant social media, Internet rumors, as an important source of market information and an exogenous financial risk, can lead to the collapse of investor confidence and the further propagation of financial risks, which can damage the financial system and even lead to social unrest. With additional availability of computing techniques, we attempt to uncover the media information effects in the stock market and seek to provide researchers with 1) a theoretical reference for a comprehensive understanding of such a complex network, 2) accurate prediction of future data, and 3) design of efficient and reliable risk intervention models. Based on the data of China's Stock Market, this study uses machine learning to investigate social media rumors to reveal the interplay of social media rumors and stock market volatility. In this work, we find patterns from social media rumors from financial forums using machine learning, quantify social media rumors based on statistics, and analyze the mechanism of propagation and influence of social media rumors on stock market volatility using econometric models. The empirical results show that rumors play an important information transmission effect on stock market volatility and the constructed Internet Financial Forum Rumor Index is helpful to sense the potential impact of rumors, i.e., a significant lagged negative effect. These findings are of guidance for the optimization of the information environment, and can serve to promote the healthy and stable development of the stock market.

KEYWORDS

complex network, machine learning, stock market, social media rumors, information dissemination

1 Introduction

The impact of Internet social media on the stock market is a double-edged sword. On the one hand, its ability to disseminate information widely and freely is conducive to reducing information asymmetry among market participants, improving the effectiveness of the stock market, and maintaining the stability of the financial market. On the other hand, the circulation of irregular, one-sided or unconfirmed information tends to impact stock prices, mislead investors, and can seriously affect the confidence of market participants in the transparency and truthfulness of market information, resulting in a decrease in the financing capacity of the stock market and a misallocation of social resources. For instance, the circuit breaker mechanism[1] in China's stock market has been triggered by reports on "devaluation of RMB exchange rate", "imminent release of many restricted sales", "geopolitical instability", "overseas transfer of domestic capital", and "the unstable situation". These rumors spread rapidly through the Internet social media, leading to weakened investor confidence and collective position reduction, financial risks, and even social unrest. Therefore, in the era of big data, it has become an important and urgent challenge in today's world to study the mechanism of the inherent influence of rumors on stock market volatility, capitalizing on the massive information on the Internet to ensure the sound operation of financial systems [1–3].

At present, there are two deficiencies in the research on the impact of rumors on the stock market [4–11]. First, we are yet to find any establishment of a single scalar value to quantify not only the degree of rumors in online social networks but also the general performance of the stock market. Second, although econometric regression models have been widely used to study the influence of rumors on stock market, such studies only focus on the one-way effect without considering how the stock market performance may feedback to online social networks. In view of the first deficiency, the Internet Financial Forum Rumor Index (IFFRI) is constructed based on the relative number of Eastmoney[2] forum rumors in time variation and spatial comparison according to the generalized attribute of statistical index. IFFRI is used to indicate the relative number of changes in social media rumors. It expresses the comprehensiveness and variation of rumors. Aiming at the second deficiency, IFFRI is added into the GARCH model to solve the problem of single investor sentiment variable in existing studies. When IFFRI is added, the intermediate variable reflects both rumor characteristics and stock market characteristics. This makes this paper better analyze the fluctuation law of social media rumors' influence on the stock market. Specifically, the contribution of this study is twofold. At the data level, we have collected and identified 430,424 rumors in China's Stock Market, which are of great value for further exploration of related issues supported by these sufficient data. At the rumor research level, this study provides a new perspective and opportunity to improve the effectiveness of stock market information disclosure, and plays a positive role in promoting the healthy and stable development of the capital market.

In this paper, we first review relevant research literature. Then, we present a social media rumors detection method, IFFRI generation principle, and empirical research model of rumor's impact on stock market volatility. Last, we design experiments to examine their effects and analyze ramifications from these experimental results to reveal the interaction between the informational and financial spaces.

2 Related work

2.1 Specialized social media on financial markets

Rose was the first researcher who proposed the impact of rumors on the stock market. By analyzing samples collected by hand over 2 years, he found that rumors can have a short-term impact on stock prices, leading investors to buy and sell [12]. Later on, similar studies have been increasingly conducted to explore the effects of social media on the stock market. In particular, Diefenback manually searched for each unsubstantiated rumor that appeared in the Wall Street Journal's market Rumors section [13]. Davies and Canes analyze the "Market Rumors" section of the Wall Street Journal and find that positive rumors have a positive impact on stock prices, while negative rumors have a negative impact. Pound et al. found through newspapers and magazines that rumors of corporate mergers and acquisitions had little impact on market fluctuations, and there was arbitrage behavior before rumors were announced [14]. Huth et al. found through media news that rumors have more impact on large-scale enterprises [15]. Barber et al. filter rumors by manually reading the "rumors" section of Business Week [16]. Kiymaz et al. detect rumors by analyzing stock market rumors in the Turkish media one by one, and found that rumors in the categories of "earnings" and "foreign takeover" had a more significant impact on stock market

1 The circuit breaker mechanism is a mechanism that sets a melting price for a contract before it reaches a stop, so that contract buying and selling quotes can only be traded within this price range for a period of time.

2 Eastmoney is one of the most visited and influential financial and securities portals in China, and has always been in a leading position among financial and economic websites in China.

volatility by collating media information [17]. Clarkson et al. investigated the relationship between rumor and abnormal returns by manually selecting rumor posts as rumor events [18]. Spiegel et al. manually selected the rumors on Israeli Internet forums and found that the rumors confirmed the significant abnormal returns in the stock market in the first 5 days [19]. Zhao et al. used false or misleading information publicly published in official media and clarified by listed companies as the study sample [5]. However, most studies mainly relied on manual identification of rumors, which is time consuming and with human bias. Some researchers have taken a further step by utilizing machine learning algorithms to identify rumors. For example, Li carried out research on network rumor recognition based on Naive Bayes classification [20]. Liu researched the detection technology of microblog rumors of unexpected events based on machine learning [21].

This paper uses machine learning method to realize the detection of social media rumors in stock market. Based on the results, we analyze the performance characteristics of social media rumors in Chinese stock market, such as sentiment polarity, time, sector and inter-industry and so on.

2.2 Sentiment polarity classification

In the study of sentiment classification in financial media, Das et al. mine investor sentiment from stock message boards and compare the efficiency of different classifiers. Their research method has a significant effect on noise removal, and they try to apply their method to different language fields [22]. Zhu used naive Bayes classification algorithm to classify six million posts into "positive", "neutral" and "negative" categories by emotion and constructed a sentiment index and opinion dispersion index [23]. Chen et al. used the evaluation theory to classify the emotional words and behavioral words in the stock market and obtained the emotional polarity of stock news by using the statistics of financial lexicon [24]. Xu et al. used support vector machine (SVM), Bayes classifier and Rough Set Theory to predict industry and individual stock news respectively and introduced a theory of sentiment classification evaluation [25]. Meng et al. obtained a keyword lexicon of investor sentiment in China [26]. Yin et al. constructed the emotional characteristics of users and microblogs by conducting emotional analysis on the rumor texts of detection microblogs and users' historical microblogs [27]. These methods provide a promising solution for sentiment classification.

In fact, research on emotional classification based on financial media has been paid more and more attention. Nowadays, network forum becomes an important form of social media, and the influence of its information dissemination has undergone great changes from breadth to depth. Most of its manifestations are semi-structured or unstructured text, such as stock forum. When studying forum information, accurate sentiment classification is the basis of quantitative impact analysis of stock market. Based on the Chinese Financial sentiment Lexicography [67] and machine learning, we will classify the rumors in the forum of Eastmoney. We use supervised learning to extract emotion information and use evaluation indexes of Chinese emotion analysis technology to evaluate the performance of emotion classification. This paper obtains more accurate emotional polarity, which provides an important basis for capturing the relationship between rumor and stock market volatility.

2.3 Quantification of social media texts

We believe that the quantification of media information is essentially the quantification of investor sentiment based on rumors. The measurement of investor sentiment is the basic work of studying the influence of media on stock market. After obtaining the rumor text and its emotional polarity, it is necessary to conduct data standardization processing first, quantify investor sentiment, and make data preparation for studying the impact of rumor on the stock market. However, because investors are affected by subjective factors such as physiology and psychology, as well as objective factors such as social environment and macroeconomy, the quantification of investor sentiment has always been a difficult problem in academia. So far there has not been a completely ideal unified measurement method.

The index of investor sentiment can be divided into market level and company level [28]. At the market level, researchers use investor sentiment factors through investor survey, closed-end fund discount, IPO offering and first-day return, market trading volume, principal component analysis, least square, HAR-RV GAS and other sentiment measurement methods [29–40]. At the firm level, researchers measure investor sentiment by discretionary accruals, decomposed Tobin, Momentum Index, market-to-book ratios, and deviation of analyst earnings forecasts [41–49].

At present, investor sentiment has not been studied by combining market and company. The main reason is the large difference of individual investors [50]. Individual investors have a greater degree of irrationality than institutional investors [51]. The biggest characteristic of China's stock market is the majority of individual investors. With the rapid development of social media in China, investors participate in discussions and express their opinions through forums, which has become a "window" for Chinese shareholders to express their emotions. At this time, rumors gather in the forum, and through investors' reading and reprinting, the

influence of rumors is expanded, and investors' emotions and behaviors are triggered. This study will use forum information to quantify investors' subjective judgment on the market and companies, discover investor sentiment hidden in rumors, and provide a new method for extracting and measuring investor sentiment.

2.4 Effect of rumors on stock market volatility

Shiller and Le Roy were among the first to discover the "volatility puzzle" of stock returns [52, 53]. Later, more and more researchers observed the influence of investor sentiment on stock market volatility. For example, Browna et al. find that investor sentiment is positively correlated with stock market returns at weekly frequencies and vice versa at monthly frequencies [54]. Wang et al. find that changes in investor sentiment have a significant impact on Shanghai and Shenzhen stock markets returns and have an inverse correction effect on the volatility of the two markets [55]. Arindam study finds that stock returns are determined by trader sentiment on the day and investor sentiment can explain stock market return volatility [56]. Clarkson et al. found that rumors react quickly to the stock market after 10 min of appearing in online forums, realizing as abnormal returns and volume [57]. Verma finds that both individual and institutional investors' sentiments have a negative impact on stock market volatility [58]. Tetlock empirically demonstrates that pessimistic media coverage predicts downward pressure on stock market prices [59]. Kaniel et al. argue that investor sentiment has an inverse relationship with short-term stock returns [60]. Patrick argues that investor sentiment changes investors' risk aversion and has a seasonal impact on the stock market [61]. Sabherwald et al. state that online forum investor sentiment has a negative impact on next-day stock returns and volatility [62]. Antonios et al. found that investor sentiment is idiosyncratic and positively associated with abnormal returns on tender announcements during corporate takeovers, a result that goes beyond previous scholarly research on the relationship between investor sentiment and the stock market [63]. Chi et al. studied the relationship between investor sentiment and ACSI information mispricing based on CAPM, Fama-French three-factor model, and Carhart four-factor model, and found that negative investor sentiment causes asset prices to deviate from value, confirming that customer satisfaction is a valuable intangible asset in capital markets [64]. Demetrios et al. studied the impact of investor sentiment in the Greek stock market, and mid-sized stocks were most significantly affected by investor sentiment [65]. Woan-lih incorporates investor sentiment factors into the Carhart four-factor model. He found that companies that were sensitive to investor sentiment earned more outlier returns in stock repurchase. At the same time, information asymmetry will exacerbate investor sentiment and lead to a greater degree of asset mispricing [66]. These studies unveil that investor sentiment has a significant impact on stock market fluctuations, but there are few researches based on social media, especially on spreading rumors.

This paper studies the influence of rumors on stock market volatility through investor sentiment reflected in social media rumors, which is of great practical significance. In particular, we explore such rumor influence according to the different stages of rumor spreading, including generation, evaluation, and dissemination.

3 Methods

In this section, we first crawl forum data (Section 3.1) and use machine learning to detect rumors (Section 3.2). Next in Section 3.3, we define a scalar value for all rumors of each day to quantify their collective significance and positivity. Such an index is used to construct the trend evolution over time. Last, in Section 3.4, we use the GARCH model to quantitatively analyze the transmission mechanism of rumors on stock market volatility.

3.1 Data acquisition

In recent years, China's stock market experienced a major stock market crash caused by information asymmetry in 2016. Within the first four trading days of that year, the Chinese stock market triggered the circuit breaker mechanism twice and closed early, setting a precedent for the world stock market. This event is providing important data for our study, and We try to discover the mechanism of propagation dynamics of social media rumors on the stock market. Therefore, this paper focuses the sample data on the Chinese stock market from 2015 to 2016. The stock data used in this paper are obtained from the China RESSET[3] database, specifically using individual stock opening price, closing price, high price, low price, turnover rate, volatility, number of shares traded, amount traded, basic information of individual stocks, the Shanghai Composite Index (SSE) and Shenzhen Component Index (SZI).

3 RESSET (Beijing Juyuan RuiSi Data Technology Co., Ltd): The company's main business is to provide financial data services, which is one of the important databases of interest to financial practitioners, researchers and investors in China.

This paper uses the text of Eastmoney stock bar as the base data. The Eastmoney stork bar is one of the most important stock forums in China: 1) According to ALEXA[4], Eastmoney ranks among the top ten Chinese websites in the world and the first financial website in China, with a daily average page view of over 100 million, which is far ahead of domestic financial websites; 2) According i-research[5], the average monthly coverage of user visits of Eastmoney reaches 63.57 million people, occupying half of the industry, and has developed into one of the most successful Internet platforms and financial data platforms in China. It providing researchers with a massive data base; 3) Eastmoney stock bar provides an interactive exchange platform for investors, and is one of the largest financial interactive platforms in China in terms of user volume. According to the data from Ariadne, Stock Bar ranks first in China in terms of user visits and user stickiness. It posting time is accurate and high, with "seconds" as the timing unit, more accurate, and the forum data is more complete. Therefore, this paper of Chinese stock market forum rumors through the Eastmoney stock bar has considerable representativeness and reference value.

We start from the URL of Eastmoney, get the initial list of pages, and keep crawling new URLs from the current page until the URL is empty or meets the crawl termination condition. We used a train crawler (an open-source web crawler) with fast crawling speed and high accuracy for adaptive modulation to meet the crawling requirements. We collected stock posts in stock bar from 2015 to 2016, totaling about 37.8 million, with a quantity of 10 GB and a data accuracy of seconds, and the content of the crawl included: stock code, posting IP address, posting title, post content, crawl URL, reading volume, following volume, and posting time.

We pre-process the data as follows. Step 1: We write the collected data into the MYSQL database, and then export the text in the database according to the information in each piece of data and export the data of the same stock into a CSV file, so as to realize the classification by stock code. Step 2: We use the program to eliminate "distorted information, mis recorded and inappropriate samples" in the database, such as very small (less than 4 kb) or very large (more than 100 kb) text, zero reading volume, long-term suspension stocks, etc. Through the pre-processing of the crawl information, 200,000 noisy posts were eliminated, and 37.6 million stock posts were made.

3.2 Rumor detection

After obtaining the forum text, we follow the route of "text representation - feature generation - feature extraction - text classification" to distinguish "rumor" and "non-rumor".

(1) Text representation. In this paper, let the information of the share Eastmoney stock bar be D and the weight of the feature term be W_k, that is, if there is a feature term $T_i$in a forum information D_i, the vector of feature terms of $T_i$is expressed as 1, otherwise it is 0. The size of the correlation between the content of two documents is measured by the distance between the vector document vectors, which is generally calculated using the inner product or the cosine of the angle, the smaller the angle the higher the similarity (Eq. 1).

$$Sim(D1, D2) = \cos\theta = \frac{\sum_{k=1}^{n} w_{1k} \times w_{2k}}{\sqrt{\left(\sum_{k=1}^{n} {w_{1k}}^{2}\right) \times \left(\sum_{k=1}^{n} {w_{2k}}^{2}\right)}} \quad (1)$$

(2) Feature Extraction. Based on the characteristics of rumor in stock market, five feature sets are used in this paper. F_a: post content features (word frequency features, word nature features, sentiment word features, etc.); F_b: followers features (number of followers, word frequency features, word nature features, sentiment word features, etc.); F_c: publisher behavior features (number of posters' posts, number of their followers, etc.); F_d: information credibility features (website credibility, posting time period, original post or repost, authority of posters, etc.); F_e: stock market features (stock index change rate, price change of corresponding stocks, volume change and turnover rate in the time period before and after posting, etc.).

(3) Feature weight calculation. Based on the *TF-IDF* model, we added a lexical weight determination method and calculated the text feature weights based on the set of stock bar information features [67]. We form a comprehensive weight for each word item of each document in the rumor sample, from which we judge that if a word item appears in a rumor sample with high frequency and the number of texts containing it in the whole rumor sample set is small, then it has a high *TF-IDF* (Eq. 2).

4 Alexa ranking refers to the world ranking of websites, mainly divided into comprehensive ranking and classification ranking, Alexa provides a number of evaluation index information including comprehensive ranking, arrival ranking, page visit ranking, etc. Most people take it as the current more authoritative evaluation index of website visits.

5 i-research is the leading brand in China's new economy and industry digital insight research and consulting services, providing professional industry analysis, data insight, market research, strategic consulting and digital solutions to help clients improve their cognitive level, profitability and overall competitiveness.

$$TF\text{-}IDF_{i,j} = tf_{i,j} \times idf_i \quad (2)$$

The weight of lexical item i in the rumor sample $j$is noted as $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$, If the fewer documents containing lexical item i in the rumor sample, it means that lexical item i has good category differentiation ability, which will be noted as $IDF_{i,j}$, and the larger its idf will be, noted as $idf_i = log\frac{N}{df_i}$. To deal with the situation when word$i$ does not exist in the set of rumor sample and the denominator df_iin the formula is 0, the above formula is modified as follows $idf_i = log\frac{N}{df_i+1}$.

(4) Text Classification. In this paper, we use Support Vector Machine (SVM) as a classifier for rumor recognition of stock bar text messages. This paper directly uses the most representative *LIBSVM package*[6], *LIBSVM-2.88 JAVA* program to build the SVM classifier.

To check the accuracy of SVM classifier in classifying rumors on the Internet, after the SVM finished the sample test, this paper used "*Precision (P) + Recall (R) + F-measure (F)*" to ensure the accuracy of classification of all samples. $P = \frac{a}{a+b}$, $R = \frac{a}{a+c}$ (a is the number of rumors correctly classified as "rumor", b is the number of "non-rumor" incorrectly classified as rumors, c is the number of rumors incorrectly classified as "non-rumor"), $F = \frac{2 \times P \times R}{P+R}$.

(5) Emotional polarity judgment. Due to the lack of a Chinese financial sentiment lexicon, most researchers in the face of the Chinese stock market have had to use manual reading discrimination methods in order to improve the accuracy of sentiment analysis, which greatly limits the sample size and increases the subjective variability of judgment results.

We design an unsupervised sentiment analysis algorithm for a large number of documents based on the Chinese Financial Sentiment Thesaurus (CFST) [67]. Step 1, the sentiment tendency of rumor is determined using CFST and combined with syntactic analysis, and the sentences with the determined tendency are used as the training set to train a sentiment determination model for rumor sentences using SVM. Step 2, the sentences in the quasi-training utterances that are larger than a certain threshold are proposed by the SVM to determine the positive and negative polarity, and are used as new training utterances to improve the SVM. Step 3, for a new rumor, the trained SVM is used to determine the positive and negative sentiment polarity of the statements in the text, and the sentiment polarity of the whole rumor message is determined based on the sentiment polarity of the sentences in the document and the importance of the position of the sentences in the text. The advantage of this method is that it uses machine learning algorithms to overcome the low recall rate of sentiment determination based on sentiment dictionaries alone. At the same time, it avoids the time-consuming of constructing training samples manually on a large scale and is suitable for mass text processing.

(6) Validation. In this paper, the experiment is evaluated using 10-fold cross-validation.

Step 1: Find any 10,000 rumor samples (A), and then find any 10,000 non-rumor samples (B), totaling 20,000 forum text messages, as the experimental data samples. Step 2: Build the training set. From the sample data, 90% rumor samples (C) and 90% non-rumor samples (D) are randomly selected as the training set and divided into 10 copies equally as the training set. Step 3: Construct the test set. The remaining 10% of rumor samples ($E = A - C$) and 10% of non-rumor samples ($F = B - D$) are used as the test set and divided into 10 copies on average as the test set.

Finally, we obtained the performance evaluation metrics of SVM classifier: *Precision = 76.82%, Recall = 72.32%, F-measure = 74.50%*. Therefore, we consider that: 1) the classifier has good performance in classifying rumors, and can detect rumor in stock bar forums crawled by the crawler; 2) the average accuracy of the SVM classifier in classifying the emotional polarity of rumors reaches *71.5%*, which can be considered as good performance in classifying emotional polarity. The trained SVM classifier can be used to discriminate the emotion of rumors.

3.3 Rumor quantification

Based on the social media rumors data obtained in Section 3.2, this paper further constructs the Internet Financial Forum Rumor Index (IFFRI) to comprehensively measure the degree and direction of social media rumors changes, which provides the prerequisites for analyzing the mechanism of social media rumors' impact on stock market volatility in Section 4 of this paper. IFFRI is the relative number of rumors in terms of time variation and spatial comparison, which conforms to the generalized attributes of statistical indexes.

3.3.1 IFFRI

To consider the comprehensiveness and variation of rumors, IFFRI is constructed according to the different stages of rumor spreading. In fact, the characteristics of social media rumors are

6 LIBSVM package: The strong principle, high efficiency, and easy operation SVM identification and regression package developed and designed by Professor Chih-Jen Lin of National Taiwan University. The package provides open-source code, which has been compiled to be executable in WINDOWS system environment. It provides tested default parameters, and the applicant has less parameter adjustment for SVM algorithm design, and provides common kernel functions such as linear and polynomial for selection, which can easily solve specific problems in SVM algorithm.

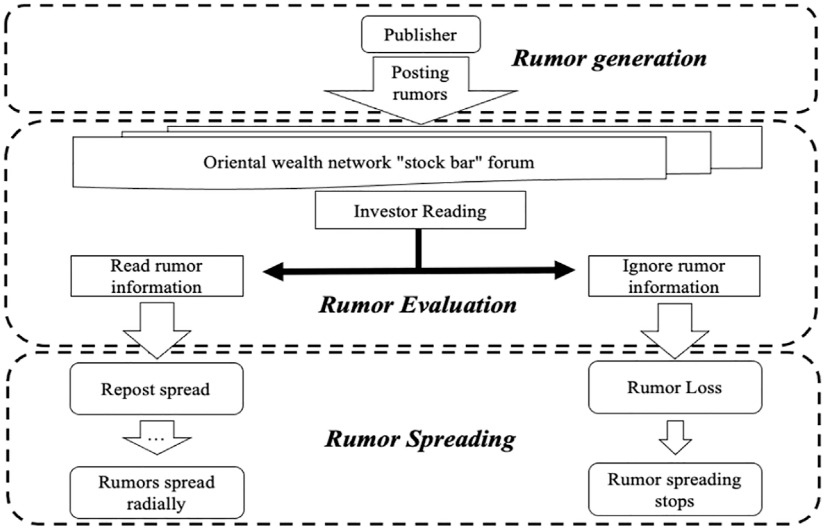

FIGURE 1
Rumor spreading behavior diagram.

essentially the same as those of traditional rumors, and their spread goes through three stages: generation, evaluation, and dissemination [68]. Rumor spreading behavior of stock forum readers is shown below (Figure 1).

1) **Rumor generation.** As the publisher's subjective evaluation and independent judgment of a listed company, rumor reflects investors' uncertainty and anxiety, which are the basis for the breeding of rumors. The publisher spontaneously generates and disseminates the information through forums, and audience receives it independently, which is typical of spontaneity and originality. The rumor at this time is "original", without interference from outside, and most truly reflects the psychological and emotional characteristics of investors. Therefore, Rumor in a forum can reflect publishers' views about companies and extract the emotional polarity of investors.

2) **Rumor Evaluation.** When investors receive rumors, they determine whether to believe them based on their own judgment. When considering other people's judgment, investors often engage in herding behavior, which is irrational, although reasonable, and is brought about by the "information cascade theory" [69]. As the survey of CNNIC (China Internet Network Information Center) shows, Chinese netizens still lack the awareness of questioning the authenticity of online news, and direct forwarding of unverified news is common. The data show that 60.3% of Internet users do not verify the authenticity of information before forwarding it directly, which plays a role in promoting the spread of false news information. Therefore, the amount of reading and commenting on rumor in a forum can reflect the extent to which investors are affected by rumor, and the emotional polarity characteristics of publishers are retained and continued.

3) **Rumor Spreading.** First, in terms of forum rumor spread, when an investor reads a rumor and believes it to be a rumor, investor quickly and actively spreads it to close friends. Rumors are passed on by people who are known and trusted (also known as "opinion leaders"), whose opinions and views are highly persuasive and are spread to increase its credibility. At the same time, rumors can flow between different social groups, with the communicator maintaining horizontal relations with other members of the group and reaching different groups through other "opinion leaders". In this way, rumor flows smoothly among multiple groups [70]. After the rumor is spread, it retains the meaning and emotional polarity of the poster, and the number of forum readings is an important indicator of the breadth of its spread. Secondly, the depth of the spread of rumors from the forum. Since there are scale differences among stocks and different sizes of impact and influence on the stock market, from a statistical perspective, the factor of market value weighting of stocks of the companies involved in rumors should also be considered. Therefore, from the comprehensive consideration of the breadth and depth of

forum rumor spread, the composition of rumor indicators should include the reading volume indicator and the market value indicator of the company involved in the rumor.

Therefore, the content and reading volume (including comment volume) of rumor in online forums imply the whole process of a rumor from generation to investors' evaluation and then dissemination, reflecting the breadth and depth of rumor dissemination, that is, the extent to which rumor "infects" the audience through "radial dissemination". We propose Attention Rate (*AR*) and Market Capitalization (*MC*); at the same time, forum rumors also contain the emotional polarity of investors' judgment on information of listed companies, that is, the emotional tendency of investment decision psychology for listed companies, and we propose Sentimental Polarities (*SP*).

$IFFRI_t$ denotes the index of financial forum rumors on day t; $AR_{i,t}$ denotes the attention index of social media rumors on day t; $MC_{i,t}$ denotes the market capitalization index of companies involved in social media rumors on day t; $SP_{i,t}$ denotes the sentiment polarity index of social media rumors on day t (Eq. 3).

$$IFFRI_t = \sum_{i=1}^{n} AR_{i,t} \times MC_{i,t} \times SP_{i,t} \tag{3}$$

In this paper, the IFFRI values of each social media rumors were calculated separately for each day in the sample period, and then summed by day to obtain the daily *IFFRI* values in the sample period. *IFFRI* provides important explanatory variables for the analysis of "social media rumors and the underlying mechanism of stock market volatility" in Section 4.

3.3.2 IFFRI factors

1) **Attention Rate (*AR*).** At day t, there are n pieces of rumor. It is necessary to weigh the importance of each piece of rumor in the attention index. The reading weight of each piece of rumor is given as follows. $AR_{i,t}$ denotes the attention index of the i^{th} rumor; $ar_{i,t}$ denotes the number of readers of the i^{th} rumor on day t; $\sum_{i=1}^{n} ar_{i,t}$ denotes the total number of rumors read at day t (Eq. 4).

$$AR_{i,t} = \frac{ar_{i,t}}{\sum_{i=1}^{n} ar_{i,t}} \tag{4}$$

2) **Market Capitalization (*MC*).** This paper matches daily market capitalization data with the market capitalization of companies involved in social media rumors on day *t*. In order to weigh the severity of the impact of the rumor on the stock market for the companies involved in the rumor, the proportion of the company involved in the market value of all listed companies (set as *J*) on t day is taken as the weight coefficient. $MC_{i,t}$ denotes the market capitalization indicator of the company involved in the i^{th} rumor; $mc_{i,t}$ denotes the market value of the company involved in rumor i at day t; $\sum_{i=1}^{j} mc_{i,t}$ denotes the total market capitalization of all listed companies in the stock market (j) at day t (Eq. 5).

$$MC_{i,t} = \frac{mc_{i,t}}{\sum_{i=1}^{j} \mathrm{mc}_{i,t}} \tag{5}$$

3) **Sentimental Polarities (*SP*).** This study divides the emotional polarity of rumor into "positive rumor" and "negative rumor". When quantifying emotional polarity indicators, *1*represents "positive rumor" and*−1* represents "negative rumor" (Eq. 6).

$$SP_{i,t} = \begin{cases} 1, if = positive\,rumor \\ -1, if = negative\,rumor \end{cases} \tag{6}$$

3.4 GARCH modeling

This study investigates the degree and direction of the effect of rumors on the volatility of the stock market. If the daily return on stocks unexpectedly rises or falls, investors will increase their expectations of variance in the next period. In this regard, a GARCH model can be used to analyze fluctuations in the effect of rumors on the stock market based on the IFFRI index and daily stock return data.

The GARCH model is the most classical model proposed by Tim Bollerslev in 1986 for describing volatility [71]. The basic principle is that the residuals reflect the magnitude of the deviation of the dependent variable from the fitted value of the mean equation, and the variance of period t can be predicted by the weighted average of the constant variance (k), the predicted value of the variance of the previous q periods (h_{t-i}) and the new information of the previous period (ε_{t-i}^2), which is particularly suitable for analysis of stock market volatility. The general GARCH model can be expressed as follows:$r_t$is the daily stock return at day t (Eq. 7), ε_t is the random error term (Eq. 8), and$h_t$is the conditional variance (Eq. 9).

$$r_t = c_1 + \sum_{i=1}^{R} \varnothing_i r_{t-i} + \sum_{j=1}^{M} B_j \varepsilon_{t-j} + \varepsilon_t \tag{7}$$

$$\varepsilon_t = u_t \sqrt{h_t} \tag{8}$$

$$h_t = k + \sum_{i=1}^{q} G_i h_{t-i} + \sum_{i=1}^{p} A_i \varepsilon_{t-i}^2 \ldots\ldots \tag{9}$$

This paper investigates the extent and direction of the impact of social media rumors on the volatility of the stock market. If the daily stock returns unexpectedly rise or fall, investors will increase their expectations of the variance of

TABLE 1 ADF test result.

Variables	T	*p*	ADF result
R_t	−16.41570	0.0000	Stable
IFFRI	−5.720275	0.0000	Stable

the next day, when a GARCH model can be used. Based on the IFFRI index and daily stock return data, we analyze the volatility pattern of the impact of social media rumors on the stock market. The process for the GARCH model is described below (Eq. 10), *m*takes the value of five in this paper[7], $IFFRI_i$ is the rumor index at day *i*. The rest of the variables are consistent with the GARCH model.

$$h_t = \mathrm{k} + \sum_{i=1}^{q} G_i h_{t-i} + \sum_{i=1}^{p} A_i \varepsilon_{t-i}^2 + \sum_{i=-m}^{m} \beta_i IFFRI_i \qquad (10)$$

3.4.1 Data validation

1) **Stability test (CHANGE to enumeration).** The GARCH model requires that each variable must be a smooth time series, i.e., with a stable trend, volatility, and cross-sectional linkage, in order to prevent pseudo-regressions. Therefore, before formally model for regression analysis, ADF unit root tests are required for daily return Rt and IFFRI values. After the ADF test, there is no unit root for both daily return Rt and IFFRI values (Table 1). It can be seen that both time series of daily return Rt and IFFRI values are smooth and can be subjected to time series model.
2) **Yield autocorrelation test**. Since the premise of the ARMA model is that the dependence between variables is manifested in the continuity of the original data in time, that is, the existence of autocorrelation of daily returns is required. Therefore, an autocorrelation test is required here for the daily return data. The test results show that there is autocorrelation in daily returns, indicating that the ARMA model can be used for model.
3) **ARMA order fixing**. From the above results, it is shown that the ARMA model can be used for model analysis, and the orders of autoregression and moving average need to be further determined. Since the order of the time series should not be too high, a total of 15 models were tried with order 3 as the maximum order, and AIC, SC, and sum of residual squares were used as judgment criteria to select the optimal model. From the fixed-order results, it can be seen that the ARMA (3, 2) model corresponds to the smallest AIC and SC, so it is determined that the ARMA (3, 2) model is used for model and analysis.
4) **Residual test.** If the residual series after ARMA model is white noise[8] series, it indicates that the ARMA model is good, and if there is heteroskedasticity in the residuals, further model of the residual series by GARCH model should be considered.

Step 1: **Residual squared autocorrelation test.** After the residual squared autocorrelation test, the autocorrelation (AC) and partial autocorrelation (PAC) coefficients show that the residuals are autocorrelated.

Step 2: **Residual variance test (ARCH-LM test).** The residual difference variance test with lags of 5 and 10 order shows that there is heteroskedasticity in the residual series by the F-value, TR value, and the corresponding *p*-value.

From the above autocorrelation test and heteroskedasticity test, it is shown that there is autocorrelation in the residual squared series and heteroskedasticity in the residual series, so further GARCH model is performed on the residual series.

3.4.2 Application of GARCH

GARCH model first requires determining the lag order of ARCH and GARCH terms. In this paper, we try to use four models GARCH (1, 1), GARCH (1, 2), GARCH (2, 1) GARCH (2, 2) and use AIC, SC as model selection criteria. From the AIC and SC values of the four models, it can be seen that GARCH (2,1) has the smallest AIC and SC values and significant regression coefficients, indicating that the model works best, indicating that the GARCH model is not autocorrelated. In summary, an ARMA (3,2)-GARCH (2,1) model is used to study the effect of rumors on stock market volatility.

4 Experiments

4.1 Descriptive statistics of detected rumors

A total of more than 400,000 social media rumors were detected according to machine learning methods (Table 2), and their descriptive statistics and feature analysis are as follows.

7 Based on the existence of short-term effects of social media rumors on the Stock Market, this paper intends to investigate whether social media rumors have an effect on Stock Market returns in the range of $[t-5, t+5]$. For the purpose of the study, this paper advances the IFFRI by five periods, thus determining $m = 5$.

8 White noise: a purely random process with expectation of 0 and a constant variance.

TABLE 2 Stock Bar volume.

Year	Information (item)	Rumors (item)	Rumor ratio (%)
2015	15,640,623	195,271	1.25
2016	22,054,662	235,153	1.07
Total	37,695,285	430,424	1.14

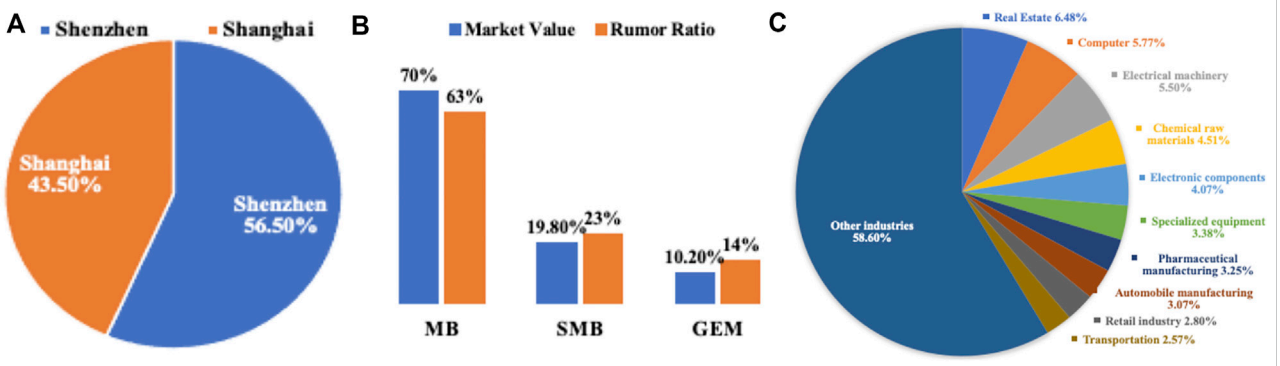

FIGURE 2
Distribution of trading venues, sectors and industry.

4.1.1 Breakdown by Trading Venues, Sectors and Industry

The stock exchange markets in China that are dominated by individual investors are Shenzhen and Shanghai. From Figure 2A, it can be seen that the Shenzhen market is significantly more than the Shanghai market. It indicates that social media rumors makers pay more attention to the Shenzhen market, which may have some relationship with the structure of China's Stock Market. Small and medium-sized stocks and GEM[9] stocks are concentrated in the Shenzhen market, and compared to the Shanghai market, there are more small and medium-sized companies and private enterprises, which are more vulnerable to the impact of social media rumors, and therefore rumor-mongers are more enthusiastic about the Shenzhen market.

When compared with the market capitalization ratio of sectors, the proportion of social media rumors involving SMB[10] and GEM is higher than that of the Main Board of Shanghai and Shenzhen (MB) (Figure 2B), indicating a higher relative concentration of SMB social media rumors in the Chinese stock market.

In terms of the distribution of social media rumors by industry, according to the 143 subcategories divided into industries by the CSRC[11], the top 10 industries in terms of social media rumors are all closely related to people's lives (Figure 2C). Among them, the three major industries, namely real estate development and operation, computer application services, and electrical machinery and equipment manufacturing, all had more than 20,000 social media rumors during the sample period, and the combined number of social media rumors for the three industries accounted for 42.9% of the top 10 social media rumors industries. Such a concentrated distribution of social media rumors reflects the high attention of media and capital to the real estate market, the use of Internet plus and intelligent manufacturing representing Industry 4.0, which is synchronized and consistent with social hot issues.

4.1.2 Temporal patterns

In terms of 1 day distribution (Figure 3A), there are two distinct peaks in social media rumors during the day: 10:00–12:00 and 14:00–15:00. It can be seen that the release of social media rumors basically coincides with the trading hours of the Chinese stock market, indicating to a certain extent that investors are

9 GEM: Growth Enterprise Market

10 SMB: Small and medium-sized board

11 CSRC: China Securities Regulatory Commission

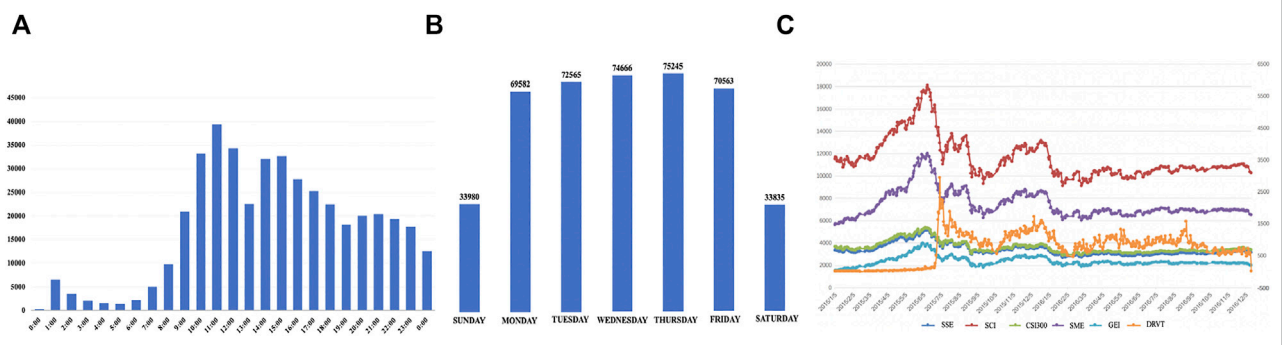

FIGURE 3
Distribution of Time. The vertical coordinate is the number of rumors, and the horizontal coordinate is the different time distribution **(A)**: 1 day; **(B)** 1 week; **(C)** observation period).

TABLE 3 AR descriptive statistics.

Statistic	Maximum	Minimum	Mean	Median	Standard deviation
AR	1	0	0.179293	0.162946	0.132082

easily influenced by the information in stock bar forums during stock trading. This is basically consistent with the Internet Report of CNNIC, which shows that "social apps have a small peak at 22:00", indicating that Chinese investors are already relying on mobile devices to a large extent. To a large extent, Chinese investors have relied on mobile devices, and it has become a habit to participate in stock market discussions through mobile APPs. From the graph, it can be seen that in addition to the two period of 10:00–12:00 and 14:00–15:00 during the day, there is also a small peak of posting at 20:00–23:00 in the evening.

According to CNNIC the Latest Internet report, "Chinese Internet users spend 28.5 h online per capita", and on which day of the week are social media rumors the most frequent? In terms of 1 week distribution (Figure 3B), we can see that the number of social media rumors gradually rises on Monday and reaches a peak on Thursday, and then drops to a slightly lower level on Friday and a trough on Saturday and Sunday. This coincides with the "Friday effect"[12] in the stock market, where rumor mongers start "warming up" on Thursday to "prepare" for trading under the "Friday effect". Public opinion preparation. Statistics show that the volume of social media rumors on national holidays (including weekends) is lower than the weekly average, indicating that Chinese investors pay the least attention to stock bar forums during holidays.

During the whole observation period (Figure 3C), the "Daily Rumor Volume Trend (DRVT)" of social media rumors basically coincides with the trend of the five stock indices: SSE, SCI, CSI300, SME and GEI[13].

4.2 Descriptive statistics of quantified rumors

4.2.1 AR

According to the AR calculation formula 1, the AR value of each social media rumors per day during the observation period was obtained as a statistical value (Table 3).

From Table 8, it can be seen that the attention level is generally small and concentrated between 0.1 and 0.2, indicating that the proportion of each social media rumors message in the total attention level of the day is between 10 and 20%. Social media rumors attention is more concentrated to reflect the extremely easy automatic alliance of online individuals with shared interests, similar interests and the same stance, driven by various factors such as information screening and filtering and the spiral of silence. As social media rumors spread interactively,

12 Friday effect: Friday usually predicts what changes in policy will occur over the weekend for 2 days and makes a move to buy or sell stocks.

13 SSE: Shanghai Composite Index; SCI: Shenzhen Component Index; CSI300: Shanghai and Shenzhen 300 Stocks index; SME: SSE SME composite; GEI: Growth Enterprise Index.

TABLE 4 MC descriptive statistics.

Statistic	Maximum	Minimum	Mean	Median	Standard deviation
MC	0.043765	0.0000088	0.0008388	0.00024916	0.002512995

TABLE 5 IFFRI descriptive statistics.

Statistic	Maximum	Minimum	Mean	Median	Standard deviation
IFFRI	0.005886	0.00001729	0.00076	0.000472	0.000781

TABLE 6 ARMA-GARCH model result (IFFRI factor is not added).

Variable	Coefficient	Std. Error	z-Statistic	Prob.
AR (1)	-0.498364***	0.035401	-14.07764	0.0000
AR (2)	-0.521428***	0.031818	-16.38769	0.0000
AR (3)	0.401244***	0.035882	11.18244	0.0000
MA (1)	0.954137***	0.006331	150.7024	0.0000
MA (2)	0.980187***	0.006682	146.6990	0.0000
Variance Equation				
C	5.31E-07**	3.66E-07	2.452029	0.0465
RESID (-1)2	0.244192***	0.058644	4.164008	0.0000
RESID (-2)2	-0.200193***	0.059398	-3.370357	0.0008
GARCH(-1)	0.953329***	0.006911	137.9479	0.0000

Note: ***0.01level significant, **0.05 level significant, *0.1level significant.

TABLE 7 ARMA-GARCH model result (IFFRI factor is added).

Variable	Coefficient	Std. Error	z-Statistic	Prob.
AR (1)	0.176209***	0.410661	11.429086	0.0000
AR (2)	0.440349***	0.301024	3.209737	0.0035
AR (3)	0.044834***	0.033055	3.077273	0.0050
MA (1)	-0.157808**	0.403087	-2.503621	0.0254
MA (2)	-0.425043**	0.290870	-2.170462	0.0439
Variance Equation				
C	-2.15E-06	6.39E-07	-3.367238	0.0008
RESID (-1)2	0.289353	0.082471	3.508521	0.0005
RESID (-2)2	-0.264836	0.081637	-3.244084	0.0012
GARCH(-1)	0.966563	0.006404	150.9424	0.0000
IFFRI (5)	0.031634	0.026267	1.204323	0.2285
IFFRI (4)	-0.046835	0.027929	-1.676922	0.0936
IFFRI (3)	0.021234	0.024839	0.854846	0.3926
IFFRI (2)	0.011335	0.030260	0.374589	0.7080
IFFRI (1)	-0.043466**	0.019225	-2.506311	0.0243
IFFRI	0.082588**	0.042403	2.488217	0.0317
IFFRI (-1)	0.141135***	0.040618	3.474665	0.0005
IFFRI (-2)	0.157343	0.032807	0.792202	0.4221
IFFRI (-3)	-0.012868	0.019596	-0.656691	0.5114
IFFRI (-4)	-0.017211	0.020295	-0.848050	0.3964
IFFRI (-5)	0.048120	0.017079	0.817472	0.4048

Note: ***0.01 level significant, **0.05 level significant, *0.1 level significant.

investors with similar attitudes, positions and judgments begin to gradually differentiate and reorganize to form cohesive subgroups. The AR value can be used to determine the concentration trend and dispersion of attention over a period of time.

4.2.2 MC

According to the MC calculation formula 2, the MC value of each social media rumors per day during the observation period was obtained as a statistical value (Table 4).

The maximum MC value is 0.043765, and the listed company involved is “PetroChina”, while the minimum MC value is 0.0000088, and the listed company involved is “Xintai Electronics”. The median value is 0.00024916 and the standard deviation is 0.002512, which indicates a large degree of dispersion. MC values are distributed as follows, and it can be seen that the companies involved in rumors are most concentrated in small and medium-sized companies with MC values less than 0.008, and there is relatively little rumor about large and very large-sized listed companies, which is basically consistent with the analysis results of 4.3.2.

4.2.3 SP

From the perspective of rumor sentiment polarity, positive rumors for 82%, while negative rumors for only 18%. The overall characteristics of rumor polarity are unevenly distributed, indicating that positive rumors predominate among stock rumors in China. The biggest reason is related to the lack of a shorting mechanism in the Chinese stock market, as rumor-mongers cannot make profits by suppressing the stock market through negative rumors, while it is easier to make profits by creating and spreading positive rumors in an attempt to raise stock prices.

TABLE 8 Robustness test result (IFFRI factor not added).

Variable	Coefficient	Std. Error	z-Statistic	Prob.
AR (1)	0.829516***	0.136773	6.064913	0.0000
MA (1)	-0.388203***	0.139910	-2.774652	0.0055
MA (2)	-0.100298**	0.080923	-2.528561	0.0152
MA (3)	-0.164216***	0.056503	-2.906347	0.0037
Variance Equation				
C	2.26E-06***	5.15E-07	4.382215	0.0000
RESID (-1)2	0.282862***	0.060494	4.675847	0.0000
RESID (-2)2	-0.239141***	0.059202	-4.039440	0.0001
GARCH(-1)	0.952541***	0.007985	119.2930	0.0000

Note: ***0.01 level significant, **0.05 level significant, *0.1 level significant.

TABLE 9 Robustness test result (IFFRI factor is added).

Variable	Coefficient	Std. Error	z-Statistic	Prob.
AR (1)	0.256550***	0.667377	7.984415	0.0000
MA (1)	0.186862**	0.669367	2.321136	0.0301
MA (2)	0.096648**	0.309043	2.373544	0.0253
MA (3)	-0.005294**	0.149115	-2.522035	0.0117
Variance Equation				
C	0.000306***	4.96E-05	6.171424	0.0000
RESID (-1)2	0.292074***	0.067003	4.359094	0.0000
RESID (-2)2	0.214970**	0.084681	2.538586	0.0111
GARCH(-1)	-0.056850*	0.146570	-1.687869	0.0881
IFFRI (5)	0.009562	0.027771	0.344313	0.7306
IFFRI (4)	-0.040098	0.024450	-1.640017	0.1010
IFFRI (3)	0.010373	0.027976	0.370774	0.7108
IFFRI (2)	-0.016124	0.033186	-0.485871	0.6271
IFFRI (1)	-0.053343***	0.007190	-7.419483	0.0000
IFFRI	0.032761	0.024523	1.335959	0.1816
IFFRI (-1)	0.045723**	0.030420	2.400840	0.0228
IFFRI (-2)	0.013860**	0.025613	2.195575	0.0416
IFFRI (-3)	-0.030547	0.018712	-1.632492	0.1026
IFFRI (-4)	0.012949	0.018256	0.709281	0.4782
IFFRI (-5)	-0.024424	0.024564	-0.994283	0.3201

Note: ***0.01 level significant, **0.05 level significant, *0.1 level significant.

4.2.4 IFFRI

The "Internet Financial Forum Rumor Index" (IFFRI) constructed in this paper consists of Attention Rate (AR), Market Capitalization (MC), and Sentimental Polarities (SP). According to the IFFRI definition (Eq. 4), the IFFRI values were calculated for each day of the sample period (Table 5).

4.3 Impact of social media rumors on stock market volatility

4.3.1 Empirical results

Step1: **Without IFFRI factor.** ARMA (3,2)-GARCH (2,1) is used for model without the IFFRI value regression results. The coefficient of variance equation variable is significant (at the 1% level) (Table 6), indicating that there is no sequence autocorrelation in the model.

Step2: **ARCH-LM test.** To test whether the ARMA (3,2)-GARCH (2,1) model eliminates the ARCH effect of the residual error sequences, ARCH-LM tests of 5 and 10 orders of lag are carried out for residual error sequences. The results show that the corresponding p values of F and the TR^2 values of 5 and 10 orders of lag are all greater than 0.1. We thus accept the null hypothesis of "the ARCH effect does not exist in residual errors"; that is, residual errors no longer have an ARCH effect, and residual information is extracted cleanly. This shows that the variance equation estimation is correct, and the model has strong explanatory power.

Step3: **Incorporation of IFFRI.** To study the influence of rumors on Stock Market volatility, the IFFRI value is added to the variance equation. The regression results show (Table 7) that rumors have a significant effect on stock volatility in $t-1$, t, $t+1$. In the $t-1$ *and* t, rumors have a positive response to stock market volatility while in $t+1$, there is a negative response. This indicates that rumors affect stock volatility in the current and next period, and both have positive response. However, as rumors may spread in advance and be fed back into stock volatility, rumors appear relatively backward, showing negative responses.

4.3.2 Robustness test

To test the robustness of the ARMA-GARCH model, we use the CSI 300 instead of the SSE Composite Index to conduct ARMA-GARCH model for all samples during the observation period. The specific processes for data preparation and model construction are similar to those mentioned above and are not repeated here. The regression results are described below.

Step 1: **IFFRI factor not added.** ARMA (1,3)-GARCH (2,1) is used for model, and the regression results for values without IFFRI are obtained (Table 8). The coefficient of variance equation variable is significant (at the 1% level), indicating that there is no sequence autocorrelation in the model.

Step 2: **ARCH-LM test is performed.** To test whether the ARMA (1,1)-GARCH (1,1) model eliminates the ARCH effect of the residual error sequence, the ARCH-LM test of the residual error sequence lagged by 5 and 10 orders is carried out. The results show that the corresponding P values of F and the TR^2 values of lag orders of 5 and 10 are all greater than 0.1. The null hypothesis of "the ARCH effect does not exist in residual error" is accepted; that is, the residual error no longer has an ARCH effect, and the residual error information is extracted cleanly. This shows that

the variance equation estimation is correct, and the model has strong explanatory power.

Step 3: **Add the IFFRI factor.** To study the influence of rumors on stock volatility based on the CSI 300 Index, the IFFRI value is added to the variance equation. The regression results show (Table 9) that rumors have a significant effect on the volatility of stocks in $t-2, t-1$, *and* $t+1$. The coefficient signs and significance of $t-1$ *and* $t+1$ remain unchanged. This is consistent with the regression results obtained based on the Shanghai Composite Index, indicating the robustness of the ARMA-GARCH model. Although $t-2$coefficient changes from insignificant to significant, the time span is close and small. Thus, we can conclude that it does not affect the empirical results of the ARMA-GARCH model.

4.3.3 Experiment summary

The ARMA-GARCH model and the variance equation regression results indicate that rumors affect stock market volatility. Specifically.

(1) The empirical results corroborate the behavioral finance perspective. In the stock investment process, a person as a system, when acquiring external information, encodes and evaluates it to form a unique investor sentiment and then makes behavioral decisions [72]. Since the information may not be complete or accurate feedback of truth, it can lead to a large amount of judgment bias in the cognitive process of investors [73]. Coupled with the fact that investors have limited cognitive resources and do not follow Bayes' law completely, they may take uncertain information as true and accurate information as long as the probability of its occurrence is higher [74]. As a result, people psychologically feel anxious when presented with unknown information and try to reduce this anxiety, leading to a level of irrationality in investor cognition that is exacerbated by the fact that people do not readily change their previously made, although suboptimal, decisions. Under the effect of investor sentiments, investors exhibit limited rationality, such as herd behavior, under- or over-reaction, and other behavioral characteristics, which in turn affect stock market volatility. The empirical results confirm that social media rumors trigger changes in investor sentiments, and that investors' behavioral decisions deviate from the optimum. Thus, rumors play an important information transmission effect on stock market volatility through investor sentiments.

(2) Consider the stock volatility at time t, and quantification of rumors *a priori*. Rumors at time $t-1$and t have a positive response to stock market volatility at t. That is, empirical results show that there is a positive effect on the stock market volatility at time t of rumors at $t-1$ and t. The greater the IFFRI at $t-1$ and t, the greater its effect on the stock market volatility at t. It owns the attitude of "rather believe it" overconfidence which coincidently matches that of House Money Effect. At the same time, we can find that if there is a positive response of rumors to stock market volatility at $t-1$, it supports the explanation that rumors have some advance effect on stock market volatility; that is, even though rumors have been spread on the Internet, investors show moderate caution about rumors and do not make decisions easily, driving investment decisions and triggering stock market volatility only on the day following the receipt of rumors, t.

(3) On the other hand, interestingly at $t+1$, rumors have a negative response to stock market volatility; i.e., have a significant lagged negative effect. In other words, the higher the stock market volatility at t, the weaker the rumor at $t+1$. It is probable that the network spammers who generated significant rumors at $t-1$ have achieved driving the market volatility at t, and then become less motivated to continue at $t+1$.

These findings suggest that social media rumor has a short-term effect on stock market volatility, which consolidates the previous studies that digital information, including news, social media, rumors, etc., has quick effects on stock market after it is released [4, 67]. In addition, the IFFRI is able to sense the potential impact of rumors on stock market movements, thus helping market regulators make more timely risk warnings and interventions to serve the promotion of healthy and stable stock market development.

5 Concluding remarks

This paper identifies a large number of stock forum rumors through machine learning methods and constructs a framework around IFFRI to quantify stock market investor sentiment. The empirical results show that the constructed IFFRI has a good ability to explain the influence trend of the Chinese stock market, and it is a comprehensive and timely index that accurately reveals the linkage of social media rumors on the stock market. This indicates that IFFRI is a suitable index for measuring investors' sentiment in the stock market. In particular, this study uses machine learning to find patterns from social media rumors and quantifies rumors based on statistical data, which sheds light on the application of large-scale market data in stock market volatility tracking, especially with the explosive growth of online data. On the application side, this study uses an econometric model to analyze the impact of rumors on the volatility of China's stock market and has a preliminary understanding of the interplay between the two, which provides a new perspective and concrete practice for the research on the transmission mechanism of social media information in the stock market.

This paper is based on a study of the Chinese stock market, and hence has two limitations. First, there is no short selling mechanism in the Chinese stock market, and profits can only be

made when stocks rise. For the same reason, as long as the house controls the majority of shares outstanding, the price is likely to move up and down in accordance with the house's will. Therefore, in China's stock rumors, rumor spammers cannot reap profits by suppressing the stock market through negative rumors; instead, they can only rely on positive rumors. Second, China's stock market implements the inter-day trading patterns, where traders cannot sell stocks purchase on the same day regardless of the degree of fluctuation in the rest of the day. In contrast, stock markets in many major economies does not have the inter-day trading patterns, which allows investors to immediately act on novel information about the stock market. This constitutes a substantial difference in the impact of rumors on stock market volatility.

This work can be extended in a few interesting ways as future research. 1) We intend to study the linkage and compound effect of rumors on the stock market. The current research mainly focuses on the direct impact of online rumors on listed companies or the impact of the entire stock market, and does not involve the more general one-to-many, many-to-one, many-to-many, or even secondary impacts of online rumors on listed companies. Therefore, we will try to use the results from the area of complex networks to study the behavior characteristics of listed companies in Internet media. By building an enterprise media relation network and analyzing the topological features of the network, we will study the overlapping effects of multiple social media rumors on the intersection companies. 2) We will conduct research on the impact of social network spammers to unveil the mechanism of rumors' influence on the stock market. At present, the research of rumors is mainly focus on general rumors identified from forum, and few involve the rumors generated by network spammers. Through the comprehensive and intelligent application to identify various rumors, we will fathom the extent of different rumors and analyze their impact on the stock market.

Through this research, we found that social media information including rumors has become an important part of external information in the stock market and even the entire financial market. How to establish a prevention mechanism with effective participation and joint supervision of all parties in the market is our next research goal. We also look forward to similar research in this vein, so as to facilitate benign interactions between financial markets and their external information environments.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

HZ, YC, RW, WR and JT designed the research, performed the research, analyzed the empirical data and wrote the paper.

Funding

This work was partially supported by The Ministry of Education of Humanities and Social Science Project of China (No.19YJA630110).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Jin BW, Zhang Q. Stock market reactions to adverse ESG disclosure via media channels. *Br Account Rev* (2022) 54(1):101045. doi:10.1016/j.bar.2021.101045)

2. Wu B, Wang S, Zeng C. Forecasting the U.S. Oil markets based on social media information during the COVID-19 pandemic. *Energy* (2021) 226:120403. doi:10.1016/j.energy.2021.120403

3. Yeşiltaş S, Şen A, Arslan B, Altuğ S. A twitter-based economic policy uncertainty index: Expert opinion and financial market dynamics in an emerging market economy. *Front Phys* (2022) 10:864207. doi:10.3389/fphy.2022.864207

4. Tetlock PC. Does public financial news resolve asymmetric information? *SSRN J* (2010) 23(9):3520–57. doi:10.2139/ssrn.1303612

5. Zhao J, He X, Wu F. Research on rumors in Chinese stock market: Spreading, refuting and impact on stock price. *Manag World* (2010) 11:38–51. doi:10.19744/j.cnki.11-1235/f.2010.11.005

6. Jiang CQ, Liang K, Chen H, Ding Y. Analyzing market performance via social media: A case study of a banking industry crisis. *Sci China Inf Sci* (2014) 57(5):1–18. doi:10.1007/s11432-013-4860-3

7. Yang SY, Yin S, Mo K, Liu A. Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance* (2015) 15(10):1637–56. doi:10.1080/14697688.2015.1071078

8. Luss R, D'Aspremont A. Predicting abnormal returns from news using text classification. *Quantitative Finance* (2015) 15(6):999–1012. doi:10.1080/14697688.2012.672762

9. Dimpfl T, Stephan J. Can Internet search queries help to predict stock market volatility? *Eur Financial Manag* (2016) 22(2):171–92. doi:10.1111/eufm.12058

10. Kumar A, Raj Sangwan S. Rumor detection using machine learning techniques on social media. *Int Conf Innovative Comput Commun* (2018) 2: 213–21. doi:10.1007/978-981-13-2354-6_23

11. Pathak AR, Mahajan A, Singh K, Patil A, Nair A. Analysis of techniques for rumor detection in social media. *Proced Comp Sci* (2020) 167:2286–96. doi:10.1016/j.procs.2020.03.281

12. Rose AM. Rumor in the stock market. *Public Opin Q* (1951) 15(3):461–86. doi:10.1086/266330

13. Diefenbach RE. How good is institutional brokerage research? *Financial Analyst J* (1972) 28(1):54+56–60.

14. Davies PL, Canes M. Stock prices and the publication of second-hand information. *J Business* (1978) 51(1):43–56. doi:10.1086/295983

15. Huth WL, Maris BA. Large and small firm stock price response to "heard on the Street" recommendations. *J Account Auditing Finance* (1992) 7(1):27–44. doi:10.1177/0148558x9200700103

16. Barber BM, Douglas L. The "dartboard" column: Second- hand information and price pressure. *J Financial Quantitative Anal* (1993) 28(2):273–84. doi:10.2307/2331290

17. Kiymaz H. The stock market rumours and stock prices: A test of price pressure and size effect in an emerging market. *Appl Financial Econ* (2002) 12(7):469–74. doi:10.1080/09603100010005852

18. Peter M, Joyce D, Tutticci I. Market reaction to takeover rumour in Internet discussion sites. *Account Finance* (2006) 46(1):31–52. doi:10.1111/j.1467-629X.2006.00160.x

19. Spiegel U, Tavor T, Joseph T. The effects of rumours on financial market efficiency. *Appl Econ Lett* (2010) 17(15):1461–4. doi:10.1080/13504850903035873

20. Li. W. Research on Internet rumor identification based on plain bayesian classification. *Comp Eng Sci* (2022) 44(03):495–501. doi:10.3969/j.issn.1007-130X.2022.03.015

21. Liu X, Chen L. Research progress of rumor identification technology for breaking news microblogs based on machine learning. *Netw Security Tech Appl* (2022) 5:54–6. doi:10.3969/j.issn.1009-6833.2022.05.031

22. Das SR, Chen MY. Yahoo! For amazon: Sentiment extraction from small talk on the web. *Manag Sci* (2007) 53(9):1375–88. doi:10.1287/mnsc.1070.0704

23. Zhu Y. The influence of network information on stock market. [PhD thesis]. Zhejiang, China: Zhejiang university (2013).

24. Chen Q, Lian W. Sentiment classification of Internet stock news based on text mining technology. *China Market* (2015) 24:234–5. doi:10.13939/j.cnki.zgsc.2015.24.234

25. Xu W, Li Y. Quantitative analysis of the impact of industry and individual stock news on stock price. *Money China* (2015) 20:31–2. doi:10.16266/j.cnki.cn11-4098/f.2015.13.025

26. Meng X, Meng X, Hu Y. Research on investor sentiment index based on text mining and baidu index. *Macroeconomics* (2016) 1:144–53. doi:10.16304/j.cnki.11-3952/f.2016.01.014

27. Yin P, Cheng P, Pan W. Early detection of microblog rumors based on integrated learning. *Microelectronics Comp* (2021) 38(01):83–8. doi:10.19304/j.cnki.issn1000-7180.2021.01.015

28. Huang H. *Investor sentiment, credit financing and corporate investment*. Beijing, China: Economic Science Press (2015).

29. Brown GW, Cliff MT. Investor sentiment and the near-term stock market. *J Empirical Finance* (2004) 11(1):1–27. doi:10.1016/j.jempfin.2002.12.001

30. Lee CMC, Shleifer A, Thaler RH. Investor sentiment and the closed-end fund puzzle. *J Finance* (1991) 46(1):75–109. doi:10.1111/j.1540-6261.1991.tb03746.x

31. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. *Proc EMNLP* (2002) 79–86. doi:10.48550/arXiv.cs/0205070

32. Pontiff J. Closed-end fund premia and returns implications for financial market equilibrium. *J Financial Econ* (1995) 37(3):341–70. doi:10.1016/0304-405X.(94)00800-G

33. Loughran T, Ritter JR, Rydqvist K. Initial public offerings: International insights. *Pacific-Basin Finance J* (1994) 2:165–99. doi:10.1016/0927-538x(94)90016-7

34. Alexander L, Nanda V, Singh R. Hot markets, investor sentiment, and IPO pricing. *The J Business* (2006) 79(4):1667–702. doi:10.1086/503644

35. José A. Scheinkman and wei xiong. Overconfidence and speculative bubbles. *J Polit Economy* (2003) 111(6):1183–219. doi:10.1086/378531

36. Cheng X, Lu J. Research on validity of investor sentiment index based on technical analysis index. *J Manag Sci* (2018) 31(01):129–48. doi:10.3969/j.issn.1672-0334.2018.01.01

37. Baker M, Stein JC. Market liquidity as A sentiment indicator. *J Financial Markets* (2004) 7(3):271–99. doi:10.1016/j.finmar.2003.11.005

38. Huang G-B, Zhou H, Ding X, Zhang R. Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst Man Cybern B* (2012) 42(2):513–29. doi:10.1109/TSMCB.2011.2168604

39. Wang Z, Hao G. Managing hypercholesterolemia and preventing cardiovascular events in elderly and younger Chinese adults: Focus on rosuvastatin. *Clin Interv Aging* (2014) 7:1–8. doi:10.2147/CIA.S41356

40. Shen Y, Yan X. Volatility and VAR prediction of CSI 300 index: Har-rv GAS model based on investor sentiment. *J Zhejiang Univ (Science Edition)* (2022) 49(1): 66–75.

41. Polk C, Paola S.The real effect of investor sentiment. Working Paper. National Bureau of Economic Research (2004). p. 105. 63. doi:10.3386/w10563

42. Goyal VK, Yamada T. Asset price shocks, financial constraints, and investment: Evidence from Japan. *J Business* (2004) 77(1):175–99. doi:10.1086/379866

43. Gao Y, Yang X, Wei X, He Y. Investor sentiment and real business performance. *Mod Business* (2022) 5:83–6. doi:10.14097/j.cnki.5392/2022.05.036

44. Zhang G, Wang M. Investor sentiment and actual investment of Chinese listed companies. *South China J Econ* (2007) 3:3–14. doi:10.3969/j.issn.1000-6249.2007.03.001

45. Jegadeesh N, Titman S. Returns to buying winners and selling losers: Implications for stock market efficiency. *J Finance* (1993) 48(1):65–91. doi:10.1111/j.1540-6261.1993.tb04702.x

46. Hua G, Zhou S, Liu Z, Jin G. Industrial policies, investor sentiment and enterprise resource allocation efficiency. *J Finance Econ* (2021) 47(01):77–93. doi:10.16538/j.cnki.jfe.20200917.303

47. Ma Y, Yang W, Jiang Y. How does investor sentiment affect a company's share price? *Finance Forum* (2020) 25(05):57–67. doi:10.16529/j.cnki.11-4613/f.2020.05.007

48. Baker M. Capital market driven corporate finance. *Annu Rev Financ Econ* (2009) 1:181–205. doi:10.1146/annurev.financial.050808.114245

49. Gilchrist S, Himmelberg CP. Gur huberman. Do stock price bubbles influence corporate investment. *J Monetary Econ* (2005) 52(4):805–27. doi:10.1016/j.jmoneco.2005.03.003

50. Huang S, wang Z. An analysis of the impact of investor sentiment on asset price: An empirical study based on Chinese stock market. *Price: Theor Pract* (2015) 11:109–11. doi:10.19851/j.cnki.cn11-1010/f.2015.11.038

51. Li G, Tang G, Liu L. Irrational linkage between stock name and stock price -- A study on China's A-share market. *J Manag World* (2011) 01:40–188. doi:10.19744/j.cnki.11-1235/f.2011.01.007

52. Shiller RJ. Do stock prices move too much to Be justified by subsequent changes in dividends? *Am Econ Rev* (1981) 71:421–36.

53. LeRoy SF, Porter RD. The present-value relation: Tests based on implied variance bounds. *Econometrica* (1981) 49(3):555–74. doi:10.2307/1911512

54. Browna GW, Cliff MT. Investor sentiment and the near-term stock market. *J Empirical Finance* (2004) 11(1):1–27. doi:10.1016/j.jempfin.2002.12.001

55. Wang M, Sun J. Chinese stock market returns, earnings volatility and investor sentiment. *Econ Res J* (2004) 10:75–83.

56. Bandopadhyaya A, Jones AL. Measuring investor sentiment in equity markets. *J Asset Manag* (2005) 7(3):208–15. doi:10.1057/palgrave.jam.2240214

57. Clarkson PM, Joyce D, Irene T. Market reaction to takeover rumour in Internet discussion sites. *Account Finance* (2006) 46(1):31–52. doi:10.1111/j.1467-629x.2006.00160.x

58. Verma R, Verma P. Noise trading and stock market volatility. *J Multinational Financial Manag* (2007) 17(3):231–43. doi:10.1016/j.mulfin.2006.10.003

59. Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *J Finance* (2007) 62(3):1139–68. doi:10.1111/j.1540-6261.2007.01232.x

60. Kaniel R, Saar G, Titman S. Individual investor trading and stock returns. *J Finance* (2008) 63(1):273–310. doi:10.1111/j.1540-6261.2008.01316.x

61. Kelly PJ, Meschke F. Sentiment and stock returns: The SAD anomaly revisited. *J Banking Finance* (2010) 34(6):1308–26. doi:10.1016/j.jbankfin.2009.11.027

62. Sabherwal S, Sarkar SK, Zhang Y. Do Internet stock message boards influence trading? Evidence from heavily discussed stocks with No fundamental news. *J Bus Finance Account* (2011) 38(9-10):1209–37. doi:10.1111/j.1468-5957.2011.02258.x

63. Danbolt J, Siganos A, Vagenas-Nanos E. Investor sentiment and bidder announcement abnormal returns. *J Corporate Finance* (2015) 33(3):164–79. doi:10.1016/j.jcorpfin.2015.06.003

64. Peng C-L, Lai K-L, Chen M-L, Wei AP. Investor sentiment, customer satisfaction and stock returns. *Eur J Marketing* (2015) 5(6):827–50. doi:10.1108/ejm-01-2014-0026

65. Demetrios G, Shah C. *Investor sentiment and stock returns: Evidence from the athens stock exchange*. Munich Personal RePEc Archive (2016). Online at: https://mpra.ub.uni-muenchen.de/71243/ (Accessed August 11, 2020).

66. Woan-lih L. Sensitivity to investor sentiment and stock performance of open market share repurchases. *J Banking Finance* (2016) 71:75–94. doi:10.1016/j.jbankfin.2016.06.003

67. Qing L, Wang T, Gong Q, Chen Y, Lin Z, Sa-kwang S, et al. Media-aware quantitative trading based on public web information. *Decis Support Syst* (2014) 61:93–105. doi:10.1016/j.dss.2014.01.013

68. DiFonzo N, Bordia P, Ralph L. Reining in rumors. *Organ Dyn* (1994) 23(1):47–62. doi:10.1016/0090-2616(94)90087-6

69. Bikhchandani S, Hirshleifer D, Welch I. A theory of fads, fashion, custom, and cultural change as informational cascades. *J Polit Economy* (1992) 100(5):992–1026. doi:10.1086/261849

70. Cai Y. *Research on rumor propagation of emergent group events in digital new media environment*. Jiangxi, China: Jiangxi People's Publishing House (2014).

71. Bollerslev T. Generalized autoregressive conditional heteroskedasticity. *J Econom* (1986) 31(3):307–27. doi:10.1016/0304-4076(86)90063-1

72. Tversky A, Kahneman D. Availability: A heuristic for judging frequency and probability. *Cogn Psychol* (1973) 5(2):207–32. doi:10.1016/0010-0285(73)90033-9

73. Akerlof GA, Yellen JL. A near-rational model of the business cycle,with wage and price inertia. *Q J Econ* (1985) 100:823–38. doi:10.2307/1882925

74. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science* (1974) 185:1124–31. doi:10.1126/science.185.4157.1124