The Influence of Investor Sentiment on Stock Volatility: Empirical Tests from the Financial Social Media Platform

Yulun Han

Abstract

The way investor sentiment engenders specific economic, social, or political ramifications represents the confluence of sociology, economics, and political science as interrelated academic disciplines. The accelerated progression of the Internet has significantly amplified the capacity for investor sentiment dissemination to exert a profound impact on economic and social dynamics. In particular, the emergence of social network platforms such as Twitter and StockTwits has enabled information transmission in virtual space to penetrate more quickly and deeply into the real world. This study analyzes whether investor sentiment on the Internet affects stock market in the real world based on StockTwits big data. The research explores the correlation between investor sentiment derived from social media platforms and stock market behavior, as both factors hold substantial importance in contemporary society. By building a time series model, I conducted an integrated type of analysis of social media data and market data. The results of this study are as follows. Firstly, based on the Granger causality test, investor sentiment within the preceding six trading days has been shown to precipitate stock market volatility, thereby indicating a robust lagging and unidirectional causal relationship between investor sentiment and stock price fluctuations. Second, based on

the result of vector auto regression model, investor sentiment of stocks in different sectors is positively or negatively associated with stock market volatility.

Keywords

Text Mining, Sentiment analysis, Granger Causality, Vector Auto-Regression

1 Introduction

From a sociological standpoint, Twitter and other social network platforms have emerged as significant instruments for examining the way individual decision-making at the micro-level of contemporary society can manifest as a social phenomenon at the macro-level, thereby offering valuable resources for conducting cross-level theoretical investigations within the realm of social sciences. The stock market plays the role of "barometer" of the national economy, which is of great significance for studying the volatility of stocks. With the development of technologies such as big data and artificial intelligence, researchers begin to focus on unstructured data that cannot be handled by traditional methods. With the rapid development of social networks, there are more and more multi-domain social media platforms emerging, therefore investors are turning their attention to financial social platforms. Unlike traditional social media platforms, such as Facebook or Twitter, the financial social platform will focus on the financial industry. Financial social platform, such as StockTwits, provides investors opportunities to share and interact with insights between investors, traders, and entrepreneurs. Because the stock market is highly uncertain and is more vulnerable to a variety of external factors, investors are more susceptible to various factors in stock trading. So, investors are

increasingly relying on public opinions to make investment decisions. However, any person can post their opinions on the platform, and the authenticity of individual investing opinions cannot be guaranteed. What's more, those public opinions may lead to herd behavior among investors by technological diffusion, and it may lead investors to make irrational decisions in stock trading. Therefore, this research will investigate whether public opinions under social media platform influence stock price fluctuations. This study examines the influence of investor sentiment, as derived from public opinions expressed on social media platforms, on stock market volatility. I hypothesize that, First, there exists a significant causal relationship between investor sentiment and stock market volatility, with sentiment serving as a leading indicator of subsequent stock price fluctuations; Second, the direction and magnitude of investor sentiment have a direct impact on stock market volatility, with positive sentiment resulting in increased risk-taking and higher volatility, while negative sentiment leads to lower volatility; Third, the relationship between investor sentiment and stock market volatility varies across different sectors, with some sectors being more sensitive to sentiment-driven fluctuations than others. These hypotheses will be tested using advanced econometric techniques such as Granger causality tests and vector auto-regression models to establish causal relationships and quantify the effects of investor sentiment on stock market volatility.

2 Literature Review

2.1 Predicting Stock Markets Using Big Data

2.1.1 Social Media Data

News about the economy overall have significant impact on daily price and can account

for up to 30% of daily price variance (Evans & Lyons, 2008). Yu (2013) conducted an investigation of both conventional media (including major newspapers, television broadcasting companies, and business magazines) and social media (such as Twitter, blogs, and forums) in relation to short-term firm stock market performances. Their findings suggest that social media exhibits a more robust association with firm stock performance as compared to conventional media sources. Market prediction which is based on the combination of natural language processing, machine learning, and behavioral economy has become the latest research field (Khadjeh Nassirtoussi et al, 2014). The study about information content on Internet stock message boards achieved good results by processing positive news using computational linguistics methods and predicting stock returns and volatility the following day (Antweiler & Frank, 2004). By using the combined decision tree (DT) and support vector machine (SVM) model, the accuracy of the prediction model improved to 60 percent.

In addition, the development of information technology makes the internet become an important way for people to obtain information. Search engine technology has become an important entrance for people to obtain information, so the index of network search can represent the change of people's focus to some extent. In the United States, Google occupies an absolute market share in search engines. In the finance industry, based on the daily internet search volume and integrated the search data which public concerned (such as economic recession, unemployment, and bankruptcy), thus they constructed the Financial and Economic Attitudes Revealed by Search (FEARS) index which can well reflect investor sentiment and successfully predict short-term return reversal (Da, 2011). The research has found that the result of the prediction indicators which included network search data was significantly improved compared the prediction indicators which

included Google Trends with those based on traditional research (Vosen & Schmidt, 2011). The authors further pointed out that search data could predict future economic activities, so it will help the government in its economic decisions.

2.1.2 Text Data and Text Mining

Text mining is widely used in various fields. Text mining includes many tasks and branches, such as text clustering, text categorization, sentiment analysis, information extraction, etc. Its purpose is to convert unstructured data that computers cannot understand into structured data that computers can understand. With the development of technology, more and more scholars begin to use text data in their research.

Zhai (2007) demonstrated that by amalgamating social media reviews, financial news, and historical price data, it is feasible to construct a comprehensive investor sentiment index. Data mining methods is feasible for analyzing quantitative and qualitative data from financial reports (Kloptchenko, 2004). The result showed that text clustering performs well in predicting technology companies. Ravi (2015) systematically summarized 161 articles and studied the tasks, approaches, and applications of sentiment analysis. By studying the method of assessing the general public's sentiments and opinions from weblogs, there is a Sentiment-PLSA model to predict product sales performance and used the ARSA model to predict movie data set (Liu, 2007). The result showed that both methods were extremely effective. Many scholars use text data to monitor fraud in financial markets. Shirata and others (2011) used the bankruptcy prediction models to predict the bankruptcy of Japanese enterprises based on annual reports. The authors applied the data of 180 companies from 1999 to 2005 (90 companies were in good operation and 90 companies were bankrupt). They made predictions by extracting specific keywords using a Classification and Regression tree model (CART).

Based on the automation text analysis method, some researchers created a dictionary from Management Discussion and Analysis Sections (MD&A) of 10-Ks (Cecchini et al, 2010). The prediction results for bankruptcy reached 83.87%, and the prediction results for fraud reached 81.97%.

2.2 Investor Sentiment

Traditional financial theory, predicated on the efficient market hypothesis, posits that individuals are rational and seek to maximize their interests throughout their lives. However, this hypothesis struggles to account for certain market behaviors that deviate from rationality. To explain a variety of market anomalies, Schöbel finds that people making decisions are influenced by various factors in real-world cases. Also, irrational factors occupy the main position in the process of making investment decisions. In the stock market, investor sentiment is an important factor affecting the stock volatility, thus there are a large number of research results in the related area.

Mehra and Sah (2002) mentioned three mechanisms for investor sentiment to affect stock price in the arbitrage market: firstly, there is the systematic fluctuation of investor sentiment; secondly, investors' assessment of risk is determined by their moods in the decision-making process; thirdly, investors believe that decisions are made based on their objective judgment, thus ignore the subjective impact of mood fluctuations. Baker and Wurgler (2003) selected the discount rate of closed-end funds, first-day IPO average return rate, dividend premium, and other factors to conduct a principal component analysis based on relevant research. Finally, they found that the constructed investor sentiment could explain the stock return.

Wysocki (1999) analyzed the 50 companies with the most mentions in the posts from

Yahoo Finance and found that the number of posts could indeed predict the stock trading volume and excess returns of the next trading day. Tumarkin and Whitelaw (2001) analyzed the statistical correlation between the abnormal posting volume and content tendency of the Raging Bull stock platform and the abnormal trading volume and abnormal returns. Antweiler and Frank (2004) generated bullish, flat, and bearish investor sentiment indexes from the posts in Yahoo Finance and found that they had the ability to predict the fluctuations of the Dow Jones Index. Das and Chen (2007) extracted the opinions of small and medium investors from a financial social media platform using semantic analysis method to generate investor sentiment index and found that the index is correlated with stock market prices. In addition to the social media platform, website news also has an impact on the stock market. For example, Liang (2005) found that the content and quantity of stock news on Yahoo Finance, Raging Bull and SmartMoney were correlated with the return rate, price fluctuation and trading volume of related stocks.

However, most of these studies were conducted on one or several platforms, especially professional platforms. Recent financial research has shifted its focus from single websites to social media platforms that represent broader investor opinion. Research on Twitter shows that, the index of collective sentiment tendency of Twitter users towards a stock in the investment field can predict the return rate of the stock (Sprenger et al., 2013; Bartov et al., 2015), and this relationship is more obvious for small sized stocks or companies. In addition, relevant studies began to correlate Twitter sentiment with the overall stock market from a macro level. The study found that the collective sentiment tendency index in Twitter could predict the rise and fall of the Dow Jones Index on the next trading day with a prediction accuracy of 87.6% (Bollen et al.,

2011). In addition, Twitter sentiment is closely related to fluctuations of trading volume in stock market (Zhang et al., 2011). More importantly, the frequency of 26 financial words in the previous one or two days was significantly correlated with the overall return rate of the US stock after-market (Mao et al., 2011). This means that investor opinion on Twitter, especially buzzwords, has the effect on the stock market.

2.3 The Effect of Investor Sentiment on Stock Volatility

Many scholars have found that investor sentiment is highly correlated with the stock market. Sayim and Rahman (2015) studied the trading sentiment of Turkish investors and Turkish ISE index, by building a VAR model and impulse response function, they found that unexpected changes in rational and irrational investors' sentiment have a significant impact on ISE returns as well as volatility of returns. Perez-Liston. Huerta et al. (2016) divided stocks into different portfolios based on large and small size and used GARCH model to investigate whether investor sentiment in Iran has a differential impact on stocks of different sizes in terms of returns and their volatility, and the results show that small-sized stocks are more affected by investor sentiment. Guo (2017) isolates the long-term trend term and the stochastic disturbance term from the investor sentiment indicator. By building an asymmetric model of volatility, Guo finally found that the long-term period trend term representing fundamental optimism and the stochastic disturbance term representing speculator sentiment the stochastic disturbance term representing speculators' sentiment enhances the leverage and anti-leverage effects of stock return volatility, respectively. Zhang (2011) found that the sentiment factor constructed by their TVP-VAR model can be used as a volatility leading indicator, which predicts better in bull markets than bear markets. Second, they find that stock market volatility is highly

when the stock market is highly volatile, the sentiment index has a greater impact on China's stock market compared to the U.S. stock market. Chen (2016) found that optimistic and pessimistic investor sentiment found that optimistic and pessimistic investor sentiment have differential effects on stock volatility, but both make stock volatility increases. They also find that financing and financing securities weaken the effect of investor sentiment on stock volatility. Sprenger (2013) empirically showed that the relationship between noisy trading and stock market volatility is unidirectional, with more noisy traders leading to higher stock market volatility. Kumari et al. (2015) used ten overall market-related sentiment proxies to market-related sentiment proxies to construct a sentiment index applicable to the dry Indian market, which was investigated through VAR-GARCH and model, and found that investor sentiment has a significant impact on stock market volatility and that past investment returns and investor sentiment have a positive or negative impact on the volatility of returns. Liu (2007) used investor sentiment with three different frequencies of monthly, weekly and daily to MARCH-MIDAS model, and found that mixed-frequency sentiment has a significant impact on the long-term volatility of stock returns have a significant impact.

3 Data and Methods

3.1 Data

For data of this research, there are two parts to the dataset, which are social media data, and stock volatility data. Target company data includes 10 target companies that are the holdings of Standard & Poor's 500 Index (SPX). The Standard & Poor's 500 Index, abbreviated as S&P 500 Index, is a stock index of 500 publicly traded companies in the

United States. This stock index is created and maintained by Standard & Poor's. All companies covered by the S&P 500 Index are listed companies traded on major U.S. exchanges, such as the New York Stock Exchange and Nasdaq. Compared to the Dow Jones index, the S&P 500 includes more companies and therefore has more diversified risk and reflects a broader range of market changes. The decision to focus on investors' sentiments for the top 10 trading stocks within the Standard & Poor's 500 Index (SPX) in predicting market volatility is based on several reasons. Firstly, these top 10 stocks represent the most popular and widely-held investments in 2021, which implies that they are likely to have a more significant impact on the overall market volatility than lesser-known or smaller-cap stocks (Barber & Odean, 2008). Concentrating on these top 10 stocks, the study aims to capture the most influential sentiment-driven price movements contributing to market-wide fluctuations. Those 10 target stocks are the top 10 stocks to invest in in 2021. In addition, those 10 target stocks are in 7 different sectors, which are communication services, consumer cyclical, consumer defensive, financial, health care, technology, and real estate. Including stocks from 7 sectors – communication services, consumer cyclical, consumer defensive, financial, health care, technology, and real estate – ensures that the study captures diverse industries and investment styles. This diversified selection helps to mitigate sector-specific biases and provides a more comprehensive understanding of how investor sentiment, as expressed on social media platforms, affects stock market volatility across different sectors (Tetlock, 2007).

Social media data was amassed for ten target companies, encompassing posts shared between January 1st, 2023 and March 1st, 2023 on StockTwits, a specialized social media platform tailored for the exchange of ideas among investors, traders, and entrepreneurs. The dataset of social media data includes user id, the text of each post,

postdate, and trading decision, such as bullish or bearish. The data size of one company is about 36,000 observations. The size of the entire companies' social media dataset is approximately 1,800,000 observations.

Stock volatility data collects historical daily data of CBOE Volatility Index (VIX) in the recent 60 trading days from Yahoo Finance. The Chicago Board Options Exchange Volatility Index (VIX) is a real-time index that reflects the market's expectation of the relative strength of near-term price changes in the Standard & Poor's 500 Index (SPX). It is the index obtained from the weighted average of the implied volatility of the index options. Because it is derived from the price of short-term expiration SPX index options, it produces a 30-day forward volatility forecast volatility, the rate of price change, which is often seen as a measure of market sentiment and, in particular, the level of fear among market participants. The index is more widely known by its ticker symbol, often referred to as the "Volatility Index". It was created by the Chicago Board Options Exchange (CBOE) and is maintained by the CBOE Global Markets. It is an important indicator for the trading and investment community because it provides a quantifiable indicator of market risk and investor sentiment (Kuepper, 2022).

I collect the social media data use a web crawler and download stock volatility data from Yahoo Finance manually. Yahoo Finance is a financial media platform owned by Yahoo. It provides financial news, and official stock’s historical data, including stock quotes, press releases, financial reports. The web crawler is an automatic computer program that collects data from World Wide Web. Web crawlers can accomplish many complicated things that traditional search engines cannot. For example, people can use a web crawler to summarize all the flight information from several websites and then write a computer program to determine the best time to buy the ticket. Compared with the

traditional API method, the traditional API method may have limitations on the requests of contents and times, but the web crawler is not subject to those limitations. In addition, this research applies Python to construct the models of web crawler and data analysis.

3.2 Methods

By browsing and reading a large number of relevant literatures, I sort out the relevant research theories and methods on the impact of investor sentiment on stock market, cognize the concept of investor sentiment, clarify the sources of channels to obtain effective investor sentiment, and understand the general impact of investor sentiment on stock market volatility. Eventually, certain improvements and innovations are made based on the previous theories.

In this study, sentiment analysis is applied to the construction of investor sentiment scores. Investor sentiment and stock volatility are applied to the Granger causality test to find whether there is a causal relationship between them. Finally, the VAR model is applied to the study of the interrelationship between investor sentiment and stock market returns to explore the dynamic interrelationship between the periods. The computational methods of this research can be divided into two main sections, which are investor sentiment analysis, and time series analysis of investor sentiment and stock volatility.

3.2.1 Sentiment Analysis

For investor sentiment analysis, the essence of unstructured data processing is to transform unstructured data which cannot be recognized by computers into digital data which can be recognized. This section applies the social media dataset in the analysis and calculates the daily sentiment score for each stock in the most recent 60 days. Text mining includes a series of contents, including sentiment analysis, text classification,

topic model, etc. This research uses sentiment analysis to compute daily sentiment scores. The first step is data preprocessing. Preprocessing text can remove the words and punctuations from the text data that do not have any sentimental components. After preprocessing data, the second step is tokenization by creating a vocabulary that stores each unique word and assigns some numeric value to each distinct word. Since machine learning algorithms cannot work on the raw text directly during language processing, feature extraction is the next step to converting text into a matrix of numerical features. Using Harvard IV-4 sentiment dictionary as frequency dictionaries to count the positive and negative frequencies of each word in the social media data. The final step is to calculate daily sentiment score for each stock by using following formulas (Kannan et al., 2016):

$$Sentiment\ Score = \frac{Freq_{pos} - Freq_{neg}}{Freq_{pos} + Freq_{neg}}.$$

3.2.2 Time Series Analysis

Following system design is proposed in this research to explore the causality and correlation between investor sentiment and stock volatility by using time series analysis.

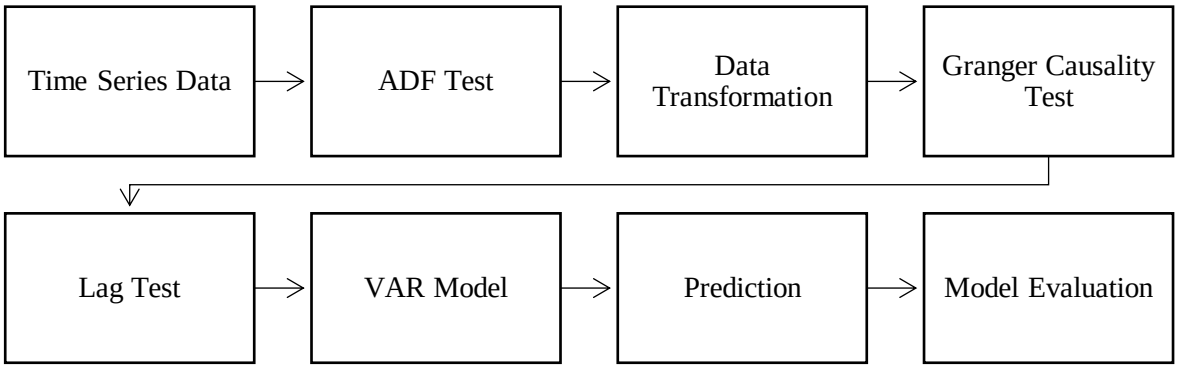

Figure 1 System Design

The time series analysis is mainly designed as the following steps. First, the data needed to be preprocessed, so I did a normality test based on the null and alternate hypothesis intuition to achieve a normal distribution for each series. Secondly, it is necessary to apply a differencing procedure to the training set in order to render the time series data stationary and smooth. However, this is an iterative process, and after our first difference, the series is already stationary. In the third step, I will conduct Granger causality test. The formal definition of Granger causality can be interpreted as whether past values of x contribute to the prediction of y, provided that the effect of past values of y on y has been explained (Maitra, 2021). If this is the case, x is said to be the Granger cause of y. Thus, the basis behind the VAR model is that every time series in the system affects each other. Granger causality tests the null hypothesis that the coefficient on past values in the regression equation is zero. Therefore, if the test yields a p-value less than the significance level of 0.05, then the null hypothesis can be safely rejected. This was performed on the original data set. Since the y terms in the equation are interrelated, y is considered as an endogenous variable rather than an exogenous predictor. To stop the problem of structural

instability, the lag length is chosen according to the AIC using the VAR framework. The next step is to fit the VAR model on training set and then used the fitted model to forecast the test set. These forecasts will be compared against the actual present in test data.

In the time series analysis of investor sentiment and stock volatility, the independent variable of investor sentiment has been one of the research hotspots. Many researchers have applied different independent variables to study the relationship between investor sentiment and stock volatility. In most cases, the higher the stock volatility, the riskier the stock trading. Therefore, this research uses sentiment score as the independent variable to analyze correlation with stock volatility as the dependent variable for each target stock by applying a regression algorithm.

4 Results

4.1 Granger Causality Test Result

In this study, the Granger causality test is employed to scrutinize the causal relationship between investor sentiment and stock market volatility. The result of Granger causality test is shown in Figure 2.

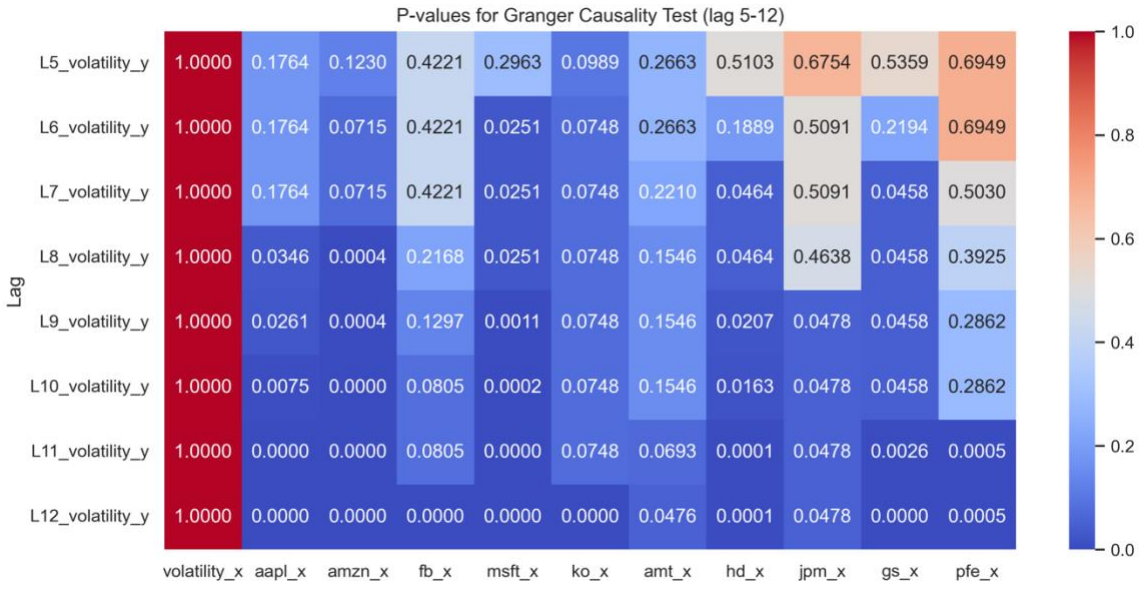

Figure 2 P-values for Granger Causality Test (lag 5-12)

Determining the optimal lag order of the time series model is not only to ensure that there is an appropriate number of lagged variables to reflect the dynamics of the model, but also to reduce the loss of degrees of freedom. If the loss of degrees of freedom is too large, the number of samples minus the degrees of freedom will leave very little, and the validity of the model is not guaranteed. Only if the degrees of freedom are sufficient, the interrelationships and interactions between variables can be better reflected dynamically.

In this study, I use AIC, SBIC, and HOIC to determine the optimal lag order. Specifically, the AIC, SBIC, and HOIC evaluation metrics all weigh the criteria for estimating model complexity and goodness of fit data but differ in the estimation and penalty terms for model goodness. I opt for the Akaike Information Criterion (AIC) to determine the optimal lag order for the VAR model. AIC is a widely recognized and frequently used criterion for model selection that balances the goodness-of-fit with the complexity of the model, ensuring that the selected model is neither underfitted nor overfitted (Akaike, 1974).

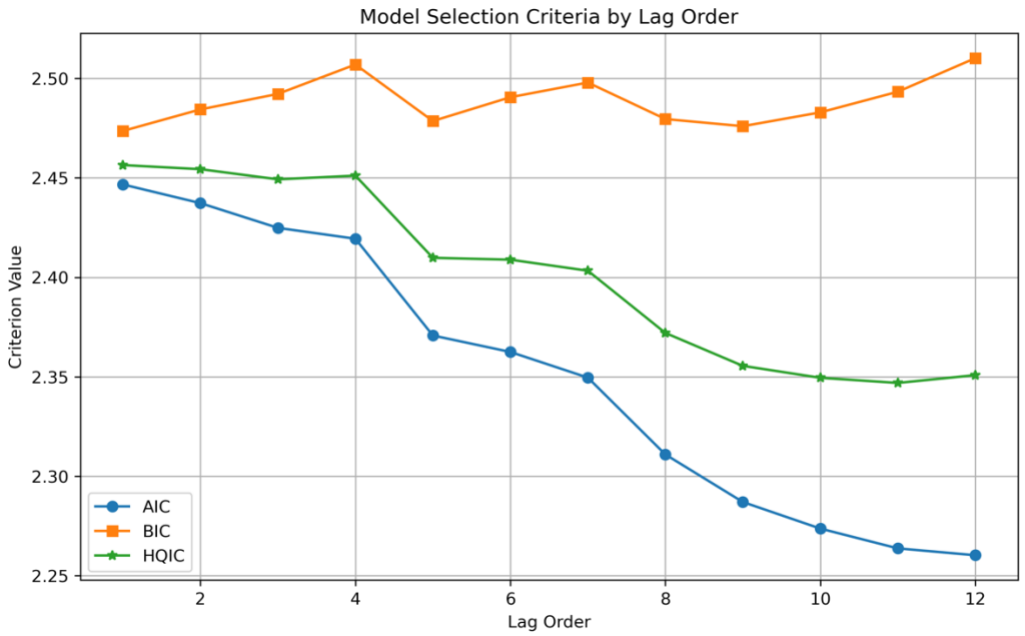

Figure 3 Model Selection Criteria by Lag Order

Upon examination of the AIC values for different lag orders, it is evident that a lag order of 6 yields the lowest AIC value of 2.36244, indicating the best model fit with minor complexity. This aligns with the general principle of selecting the model with the lowest AIC value (Burnham & Anderson, 2002). Therefore, we choose a lag order of 6 for our VAR model in this study, as it provides the most reasonable trade-off between model fit and complexity, ensuring reliable and accurate inference and forecasting.

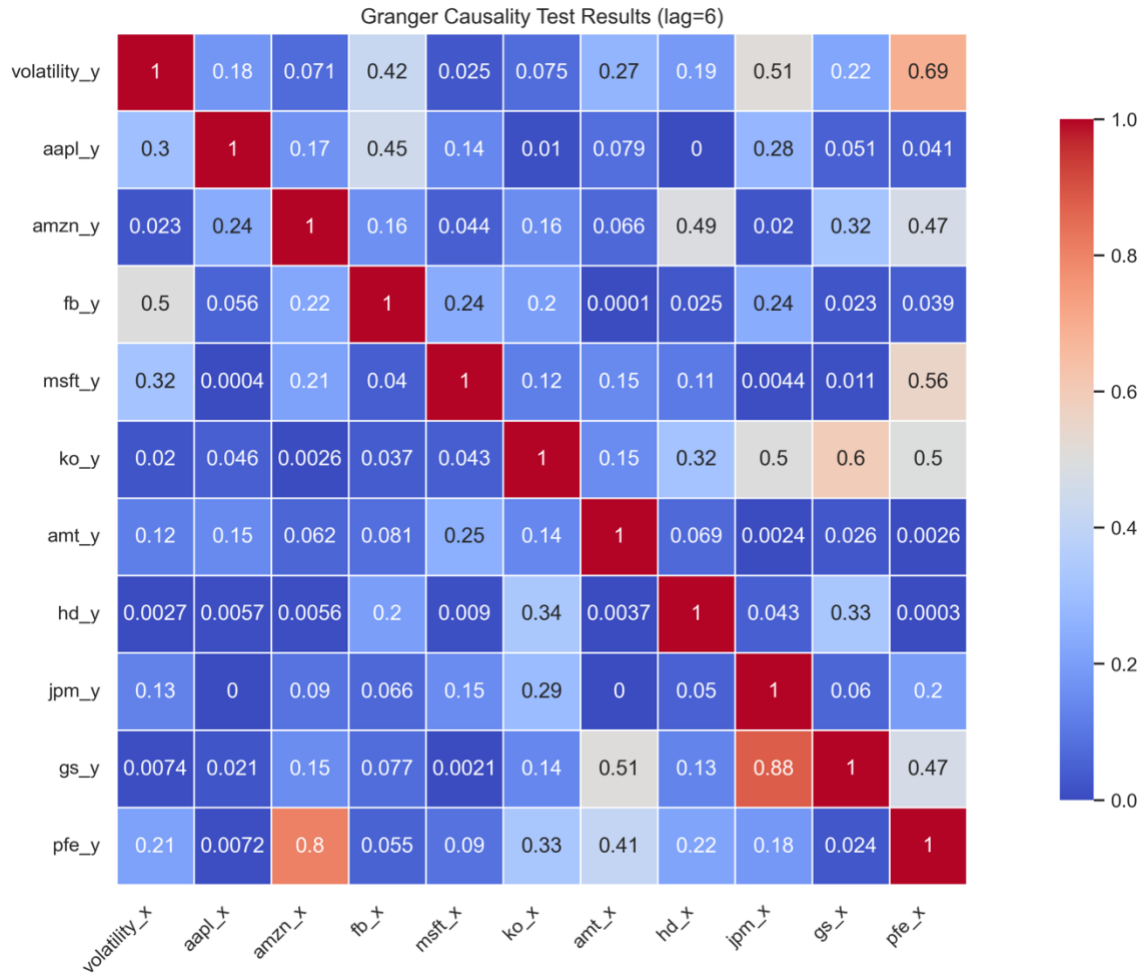

Figure 4 Granger Causality Test Results (lag = 6)

Examining the Granger causality test results reveals the p-value for the relationship between Microsoft's volatility (msft_x) and the dependent variable (volatility_y) is 0.0251. This p-value is lower than those obtained for other independent variables, such as Apple (aapl_x), Amazon (amzn_x), Facebook (fb_x), Coca-Cola (ko_x), American Tower Corporation (amt_x), Home Depot (hd_x), JPMorgan Chase (jpm_x), Goldman Sachs (gs_x), and Pfizer (pfe_x). The null hypothesis of the Granger causality test assumes no causal relationship between the variables in question. A lower p-value provides more substantial evidence against the null hypothesis, justifying the selection of 'msft' over other variables (Granger, 1969).

A commonly used significance level of 0.05 is a threshold to determine statistical significance. As the p-value for 'msft' (0.0251) falls below this threshold, the null hypothesis can be rejected, suggesting that historical changes in Microsoft's stock volatility (msft_x) have a significant Granger-causal impact on the dependent variable (volatility_y) (Hamilton, 1994).

When the lag is set to 5, investor sentiment exhibits no causality with stock volatility. However, when the lag is increased to 6, the p-value of Microsoft stock (ticker: "msft") falls below 0.05, suggesting that investor sentiment for "msft" demonstrates Granger causality with stock volatility. Furthermore, as the lag increases, more stocks exhibit causal relationships with volatility. A bidirectional Granger causality test is conducted to ascertain further the presence of a unidirectional causality (Granger, 1969).

	volatility_x	msft_x
volatility_y	1.0000	0.0251
msft_y	0.3166	1.0000

Figure 4 Granger Causality Test Results for Microsoft (lag = 6)

The bidirectional Granger causality analysis results (Figure 4) reveal that the p-values for the chi-square distribution model and the likelihood ratio test model are below 0.05. This indicates that the null hypothesis can be rejected, and the findings of this study are statistically significant. The results also confirm a correlation between investor sentiment and stock volatility, with the sixth-order lagged term of investor sentiment employed to test stock volatility in the current period. This demonstrates that the constructed investor sentiment has predictive power over stock volatility, and the model developed in this study is valid. Subsequently, a vector autoregressive model is used to construct a time series analysis model (Hamilton, 1994).

In summary, the Granger causality test results support the choice of Microsoft ('msft') as the variable with the most significant causal relationship to the dependent variable (volatility_y), given its lowest p-value among all considered independent variables. This decision aligns with methodologies proposed in previous academic research (Granger, 1969; Hamilton, 1994).

4.2 Vector Auto-Regression Model Result

The vector auto-regression model (VAR) does not require the usual regression analysis to determine the exogenous relationship, but can simultaneously study the interactions between variables, and can analyze the dynamics and interactions of variables within the system by treating the random disturbance term as an external shock. Based on these advantages, this study uses a VAR model to study the relationship between changes in investor sentiment and stock market returns.

Correlation Coefficient with Volatility	
volatility	1.000
fb	0.550
hd	0.390
aapl	0.260
amzn	0.026
gs	-0.0049
jpm	-0.011
amzn	-0.11
ko	-0.22
msft	-0.24
pfe	-0.42

Figure 5 Features Correlating with Volatility (lag = 6)

As I discovered the significant causal relation between investor sentiment and stock volatility, I want to conduct an additional time series analysis. By using the vector autoregression model, to investigate if there is a significant correlation between stock volatility and 10 stocks in the 7 sectors, such as information technology, financial, real estate, health care, etc. Based on the Figure 5, it shows “fb” (Facebook), “hd” (Home Depot), “aapl” (Apple), and “amzn” (Amazon) have positive correlation with stock volatility. In addition, “gs”, “jpm”, “amzn”, “ko”, “msft” and “pfe” have negative correlation with stock volatility.

The vector auto-regression model results display varying correlation coefficients between stock volatility and individual stocks, with some positive correlations and others negative. This phenomenon can be explained by the fact that stock prices and their

associated volatilities are influenced by many factors, including market conditions, company-specific news, investor sentiment, and macroeconomic variables (Engle, 1982; Bollerslev, 1986).

Positive correlations indicate that as the stock's price volatility increases, the stock's returns tend to increase as well. This relationship might be attributed to higher market risk, leading to increased returns to compensate investors for bearing that risk (Fama & French, 1993). In the results provided, Facebook (fb), Home Depot (hd), Apple (aapl), and Amazon (amzn) exhibit positive correlations with volatility.

5 Discussions

In the article, 10 stocks in Standard and Poor's 500 index were selected for analysis, and investor sentiment scores were constructed by using social media data from February 2022 to April 2022. Using web crawler, sentiment analysis, Granger causality, and vector autoregression model to investigate the relationship between investor sentiment and stock volatility. This empirical study shows that there is a significant correlation between investor sentiment and stock volatility. When investor sentiment fluctuates, it will affect the fluctuation of stock volatility.

The process of stock investment should be a social interaction process, including the process of investors' information immersion to the market, network interaction and emotional stress. Investors' judgment on the capital market cannot be completely independent, and investors' emotional response to market fluctuations cannot be completely rational. Individual bias will lead to group bias in the financial market, and then lead to investment or portfolio decision-making bias. In other words, investors'

market decisions are deeply influenced and interfered with each other, and this influence will be further expanded with social media public opinion's exaggeration. Especially from the perspective of information dissemination, individual investors generally lack the professional ability to collect and process market information, so they are more susceptible to the influence of public opinion.

Compared with traditional face-to-face social networks, the Internet has the advantages of low access cost, rapid time response and wide spatial links. Compared with general Internet platforms, StockTwits also has some important characteristics, such as real-time media, analytical hierarchical communication, and relative authenticity. These characteristics enable participants in the stock market, especially individual investors, to partially alleviate the problem of information asymmetry, and may also amplify the chain reaction brought by false information, thus forming the mechanism that StockTwits can profoundly influence the stock market. Therefore, in the age of fragmented network information, social media platforms have become the most important information sources for the public.

In terms of methodology, compared with traditional methods such as participative observation, deep interview and questionnaire investigation, big data research using StockTwits can better analyze specific social and economic life changes from the macro level and long-term scale. Thus, it provides a new prospect for social scientists to study the interaction between social and economic movement track, cultural behavior, and political phenomena in real life. A few years ago, Gary King (2009), the professor at Harvard University, predicted when looking into the future fifty years of social sciences that with the emergence and application of big data, the empirical basis of the whole social science research would be significantly transformed, and the breadth of research

problems would be greatly expanded with the integration of massive data. It will even accelerate the integration of qualitative and quantitative research.

This study will provide reasonable suggestions from the perspectives of both securities regulators and investors in the hope that it will help to improve investors' investment ability, reasonably guide investors' sentiment, ensure the healthy development of the capital market and better perform its barometer function.

5.1 Suggestions for Securities Regulators

A sound information disclosure system is a prerequisite for the reasonable guidance of investors' sentiment. The securities market is plagued with problems such as sluggish information disclosure and the casual dissemination of false information. Specifically, listed companies, in the case of significant information, such as major investment acts, signing of important contracts, changes in the scope of business, significant losses incurred, significant changes in the external environment, mergers, demergers, dissolutions, capital increases and reductions of companies, etc.

There are non-transparent, incomplete, untimely, and even false disclosures. Based on such low-quality information, it is difficult for investors to make correct judgments, which in turn causes stock prices to be prone to abnormal fluctuations. In addition, in today's highly developed Internet, the popularity of mobile terminals makes the spread of fake news more rapid and at a lower threshold than before, which coupled with the difficulty of judging the source of information, makes it difficult for people to discern the truth from the information they obtain and finally makes the information asymmetry in the stock market more and more serious. Information is closely related to investors' decisions, and investors' decisions will in turn have a great impact on the market. Some

wrong information may lead to a collective herding effect of investors, which in turn makes the market fluctuate drastically.

In response to the first problem, the information disclosure of listed companies should be strengthened, and the regulator should introduce corresponding laws and regulations to improve the existing information disclosure system, so that listed companies can disclose important information correctly, comprehensively, and timely, and ensure that the whole process is open and transparent. For the content of disclosure of listed companies, it is better to refine it on top of the corresponding system, listing the specific scope of public disclosure, stipulating the form of information disclosure of listed companies, the time and manner of disclosure. In general, the specific content of disclosure to reduce investor misinterpretation as a prerequisite.

In response to the second problem, regulators need to improve the information traceability and accountability system, and to strictly and severely punish the media and individuals who maliciously spread false news. At the same time, the regulation of online information needs to be strengthened, and a system that can respond to online public opinion promptly and swiftly should be established so that false and badly influenced information can be controlled in time.

5.2 Recommendations for Investors

From the investor level, investors should understand common psychological biases such as herding effect, overconfidence, representativeness bias, disposition effect, etc., and how these psychological biases affect their decision-making behavior so that they can better understand themselves and correct their unreasonable investment behavior. For example, in the second place, investors should be aware of the disadvantageous position

of individual investment in the stock market. Compared with institutional investors, individual investors lack the appropriate professional skills and expertise and are far inferior to institutional investors in terms of news. Therefore, individual investors should learn more about the corresponding investment knowledge in their daily lives, sum up more experience, develop a healthy and good investment consciousness, not to blindly trade and chase the ups and downs. Return to the essence of stock investment and enjoy the dividends of corporate development by growing with excellent companies. In addition, investors in the process of stock market trading to learn to screen information, do not blindly listen to rumors, superstitious authority and the so-called "stock recommendation experts" advice.

5.3 Strength

This article tries to apply investor sentiment from social media data to the study of stock volatility, which provides a new idea for studying the stock volatility from the micro aspect. Overall, the stock volatility is mainly related to the trend of investor sentiment. Although sentiment score can represent the social media information of individual stocks, it can only reflect part of the information. In addition, investor sentiment toward stocks in different sectors can have different levels of impact on stock volatility.

Although the time span of the study is 60 trading days, the period is relatively stable and free from large market fluctuations. The 60-day trading period examined in the study might be considered relatively stable and free from substantial market fluctuations for several reasons, including the absence of major events, favorable market conditions, and industry-specific stability. Despite this, it is crucial to recognize that such a stable 60-day period might not accurately represent long-term market behavior. Consequently,

employing an extended timeframe may prove advantageous for obtaining a more comprehensive understanding of market dynamics. Secondly, the construction process of investor sentiment indicator selection uses a single sentiment indicator that is in line with the U.S. stock market sentiment and considers the "current" and "lagged" characteristics of sentiment indicators and the possible inclusion of basic macroeconomic information in sentiment when studying related issues. The study uses sentiment analysis methods several times and constructs a more reasonable comprehensive sentiment indicator.

Currently, there is little research literature on the asymmetric effects of different investor sentiments on stock returns in the financial field. In studying the relationship between investor sentiment and volatility, this study classifies investor sentiment and uses a vector auto-regression model with dummy variables to investigate the possible asymmetric effects of investor sentiment on stock volatility under different sentiment states. Finally, in studying the relationship between investor sentiment and stock volatility, the impact of investor sentiment on stock volatility under investor sentiment shock is investigated using the investor sentiment score constructed in this study combined with a time series model.

5.4 Limitation

When constructing the investor sentiment score, I only applied the frequency dictionary to calculate each post's polarity to use as an investor sentiment score. The sentiment analysis by using frequency dictionary is too rough to get a more accurate value. Because the posts' content on social media is short and incomplete, requiring more detailed natural language processing analysis. Based on the above limitation, the data preprocessing of social media data is simple in the article, and the improvement of investor sentiment

measure also needs the continuous development of natural language processing.

In the study, only the data from social media platforms are captured and investigated, but other social data, such as news, and Google trends, are not taken into account. Compared with social media data, the data from news and Google trend are more authoritative and able to reflect more on the sentiment of the public and government. In addition, investor sentiment is an indicator that is difficult to measure. Although relevant researchers have done a lot of studies, the relationship between social media data and investor sentiment has not been verified theoretically.

There is another limitation in the research. The data cycle I used is relatively short, which is a fluctuation period within the recent 60 trading days. Therefore, the result may not be representative theoretically. In other words, what happened in 2022 does not mean it will happen in 2023. However, given the volatile nature of United States' stock market, I believe that the data analysis based on nearly 140 consecutive trading days in 6 months is relatively reliable. Incorporating a 140-day period in a stock price prediction model, as opposed to a 60-day period, can potentially enhance the reliability of the model due to several factors. These include an increased sample size that offers a more comprehensive market representation, reduced noise impact leading to the improved capture of underlying trends, consideration of seasonality and cyclical patterns, and enhanced model stability by mitigating overfitting risks. Nonetheless, it is imperative to recognize that the optimal timeframe for stock price prediction models may vary based on market conditions, stock attributes, and the specific model used. Consequently, thorough model validation and testing must be conducted to ascertain the appropriateness of the chosen time frame. On the contrary, the data in over-long period cannot guarantee the stability of time series analysis. Furthermore, it is important to acknowledge that the time series

analysis, constrained by data availability, may not demonstrate counterfactual causality in the strictest sense, despite fulfilling the criteria of the Granger causality test. Nevertheless, the investor sentiment index constructed in this study remains robust and persuasive from both theoretical and practical perspectives.

6 Conclusion

Behavioral finance came into being around 1980, when the abnormal behavior in the financial markets could not be explained by traditional finance, so the discipline of behavioral finance was created. Behavioral finance broadened the scope of traditional financial theory, overturned the assumptions of rational man and arbitrage infinity, and in the cognition of behavioral finance, it is believed that investors are finite rational, and there is almost no possibility of reaching an arbitrage free equilibrium in the market through repeated arbitrage. In behavioral finance, it is assumed that investors are finitely rational and that there is little possibility of reaching an arbitrage free equilibrium in the market through repeated arbitrage. Applying behavioral finance to real life, it can well explain various situations existing in the real market, and some unconventional behaviors can be well explained with great vitality. Although scholars have a lot of research in this area, there is also no unified systematic framework for many issues. Based on different research systems, various scholars have come up with different results. In this study, I refer to many literatures to construct a comprehensive indicator of investor sentiment, and select an indicator constructed by user comments within StockTwits, and use this indicator and stock returns to conduct Granger causality tests and vector auto-regressions to analyze the changes of investor sentiment qualitatively and quantitatively.

Finding the future trend of a stock is a crucial task, as stock trends depend on many factors. Early in the research, I hypothesized that investor sentiment and stock volatility are correlated, and that investor sentiment is likely to cause stock market volatility. Therefore, I delved into this relationship through computational methods and concluded that investor sentiment six trading days ago causes stock volatility, and that investor sentiment is correlated with stock volatility, but the correlation varies based on stock sector.

In conclusion, investor sentiment has a significant impact on stock volatility, and the investment sentiment added to social media data can better explain the influence of external factors on stock volatility than in previous studies. For the stock market, relevant financial departments should establish and improve the system construction of information disclosure. Promoting timely communication of information can reduce the volatility of the stock market and stabilize the stock volatility in a stable range.

7 Availability of Data and Material

All data generated or analyzed during this study are included in this published GitHub repository. For details on how to use the data and code in this study, please refer to the ReadMe file in the GitHub repository, which contains a detailed instruction of how the data and code are used. Please access GitHub repository, which contains all the data and code involved in this study, by clicking on the link below.

https://github.com/YLHan97/replication-materials-YLHan97

References:

Antweiler, Werner, and Murray Z. Frank. “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.” *The Journal of Finance*, vol. 59, no. 3, 2004, pp. 1259–94. *Crossref*, https://doi.org/10.1111/j.1540-6261.2004.00662.x.

Baker, Malcolm P., and Jeffrey A. Wurgler. “Investor Sentiment and the Cross-Section of Stock Returns.” SSRN Electronic Journal, 2003. Crossref, https://doi.org/10.2139/ssrn.464843.

Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. Review of Financial Studies, 21(2), 785-818.

Bartov, Eli, Lucile Faurel, and Partha S. Mohanram. 2015."Can Twitter Help Predict Firm-Level Earnings and Stock Returns?" Available at SSRN:http://ssrn.com/abstract=2631421.

Bollen Johan, Huina Mao, Xiao-Jun Zeng. 2011. Twitter Mood Predicts the Stock Market. Journal of Computational Scienc, 2(1): 1-8. DOI:10.1016/j.jocs.2010.12.007

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327.

Bryden John, Sebastian Funk, Vincent AA Jansen. 2013. Word Usage Mirrors Community Structure in the Online Social Network Twitter. EPJ Data Science, 2(1): 1-9. DOI:10.1140/epjds13

Cecchini, Mark, et al. “Making Words Work: Using Financial Text as a Predictor of Financial Events.” Decision Support Systems, vol. 50, no. 1, 2010, pp. 164–75.

Crossref, https://doi.org/10.1016/j.dss.2010.07.012.

Chen Yunsong, Fei Yan. 2016. Economic Performance and Public Concerns about Social Class in Twentieth-Century Books. Social Science Research, 10001(59): 34-51.

Coleman James S.. 1986. Social Theory, Social Research, a Theory of Action. American Journal of Sociology, 91(6): 1309-1335. DOI:10.1086/228423

Da, Zhi, et al. "The Sum of All FEARS: Investor Sentiment and Asset Prices." SSRN Electronic Journal, 2011. Crossref, https://doi.org/10.2139/ssrn.1509162.

DA, ZHI, et al. "In Search of Attention." The Journal of Finance, vol. 66, no. 5, 2011, pp. 1461–99. Crossref, https://doi.org/10.1111/j.1540-6261.2011.01679.x.

Das Sanjiv R., Chen Mike Y.. 2007. Yahoo! for Amazon:Sentiment Extraction from Small Talk on the Web. Management Science, 53(9): 1375-1388. DOI:10.1287/mnsc.1070.0704

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica, 50(4), 987-1007.

Evans, Martin D. D., and Richard K. Lyons. "How Is Macro News Transmitted to Exchange Rates?" Journal of Financial Economics, vol. 88, no. 1, 2008, pp. 26–50. Crossref, https://doi.org/10.1016/j.jfineco.2007.06.001.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3-56.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica, 37(3), 424-438.

Guo, Daiyuzhu. 2017." Investor Sentiment and Volatility Asymmetry".[J], 2017, pp. 43-52

Hamilton, J. D. (1994). Time Series Analysis. Princeton University Press.

Kannan, S., et al. "Big Data Analytics for Social Media." Big Data, 2016, pp. 63–94.,

https://doi.org/10.1016/b978-0-12-805394-2.00003-9.

Khadjeh Nassirtoussi, Arman, et al. “Text Mining for Market Prediction: A Systematic Review.” Expert Systems with Applications, vol. 41, no. 16, 2014, pp. 7653–70. Crossref, https://doi.org/10.1016/j.eswa.2014.06.009.

King, Gary. 2009."The Changing Evidence Base of Social Science Research." In The Future of Political Science:100 Perspectives, edited by Gary King, K. L. Schlozman and N.Nie.New York, NY:Routledge:91-93.

Kloptchenko, Antonina, et al. “Combining Data and Text Mining Techniques for Analysing Financial Reports.” Intelligent Systems in Accounting, Finance & Management, vol. 12, no. 1, 2004, pp. 29–41. Crossref, https://doi.org/10.1002/isaf.239.

Kuepper, Justin. “Cboe Volatility Index (VIX).” *Investopedia*, 2 May 2022, www.investopedia.com/terms/v/vix.asp.

Kumari J, Mahakud J. Does investor sentiment predict the asset volatility? Evidence from emerging stock market India[J]. Journal of Behavioral & Experimental Finance, 2015, 8: 25-39.

Liang, Xun. 2005."Impacts of Internet Stock News on Stock Markets Based on Neural Networks." In Advances in Neural Networks-ISNN 2005, edited by Jun Wang, Xiaofeng Liao and Zhang Yi. Berlin:Springer Science & Business Media:897-903.

Liu, Yang, et al. “ARSA.” Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’07, 2007. Crossref, https://doi.org/10.1145/1277741.1277845.

Mahajan, Anuj, et al. “Mining Financial News for Major Events and Their Impacts on the Market.” 2008 IEEE/WIC/ACM International Conference on Web Intelligence and

Intelligent Agent Technology, 2008. Crossref, https://doi.org/10.1109/wiiat.2008.309.

Maitra, Sarit. “Forecasting Using Granger's Causality and VAR Model.” Medium, Towards Data Science, 8 June 2021, https://towardsdatascience.com/granger-causality-and-vector-auto-regressive-model-for-time-series-forecasting-3226a64889a6.

Mao, Huina, Scott Counts, and Johan Bollen. 2011."Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data." http://arxiv.org/abs/1112.1051.

Mehra, Rajnish, and Raaj Sah. “Mood Fluctuations, Projection Bias, and Volatility of Equity Prices.” Journal of Economic Dynamics and Control, vol. 26, no. 5, 2002, pp. 869–87. Crossref, https://doi.org/10.1016/s0165-1889(01)00035-5.

Perez-Liston D, Huerta D, Haq S. Does investor sentiment impact the returns and volatility of Islamic equities?[J]. Journal of Economics and Finance, 2016, 40(3): 421-437.

Pesaran M. Hashem, Yongcheol Shin, Smith Richard J.. 2001. Bounds Testing Approaches to the Analysis of Level Relationships. Journal of Applied Econometrics, 16(3): 289-326. DOI:10.1002/(ISSN)1099-1255

Preis Tobias, Helen Susannah Moat, Stanley H. Eugene. 2013. Quantifying Trading Behavior in Financial Markets Using Google Trends. Scientific Reports(3): 1684.

Ravi, Kumar, and Vadlamani Ravi. “A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications.” Knowledge-Based Systems, vol. 89, 2015, pp. 14–46. Crossref, https://doi.org/10.1016/j.knosys.2015.06.015.

Sayim M, Rahman H. The relationship between individual investor sentiment, stock return and volatility: Evidence from the Turkish market [J]. International Journal of

Emerging Markets, 2015, 10(3): 504-520.

Scheitle Christopher P.. 2011. Google's Insights for Search:A Note Evaluating the Use of Search Engine Data in Social Research. Social Science Quarterly, 92(1): 285-295. DOI:10.1111/ssqu.2011.92.issue-1

Schöbel, Markus et al. “Social Influences in Sequential Decision Making.” PloS one vol. 11,1 e0146536. 19 Jan. 2016, doi:10.1371/journal.pone.0146536

Shirata, Cindy Yoshiko, et al. “Extracting Key Phrases as Predictors of Corporate Bankruptcy: Empirical Analysis of Annual Reports by Text Mining.” Journal of Emerging Technologies in Accounting, vol. 8, no. 1, 2011, pp. 31–44. Crossref, https://doi.org/10.2308/jeta-10182.

Sprenger Timm O., Andranik Tumasjan, Sandner Phlipp G., Welpe Isabell M.. 2013. Tweets and Ttrades:The Information Content of Stock Microblogs. European Financial Management, 20(5): 926-957.

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

Tumarkin Robert, Robert Whitelaw. 2001. News or Noise? Internet Postings and Stock Prices. Financial Analysts Journal, 57(3): 41-51. DOI:10.2469/faj.v57.n3.2449

Vosen, Simeon, and Torsten Schmidt. “Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends.” Journal of Forecasting, vol. 30, no. 6, 2011, pp. 565–78. Crossref, https://doi.org/10.1002/for.1213.

Wysocki, Peter. 1999. "Cheap Talk on the Web:The Determinants of Postings on Stock Message Boards." Working Paper, University of Michigan Business School.

Yu, Yang, et al. “The Impact of Social and Conventional Media on Firm Equity Value: A Sentiment Analysis Approach.” Decision Support Systems, vol. 55, no. 4, 2013, pp.

919–26. Crossref, https://doi.org/10.1016/j.dss.2012.12.028.

Zhai, Yuzheng, et al. “Combining News and Technical Indicators in Daily Stock Price Trends Prediction.” Advances in Neural Networks – ISNN 2007, 2007, pp. 1087–96. Crossref, https://doi.org/10.1007/978-3-540-72395-0_132.

Zhang Xue, Hauke Fuehres, Gloor Peter A.. 2011. Predicting Stock Market Indicators Through Twitter:'I hope it is not as bad as I fear'. Procedia-Social and Behavioral Sciences, 10001(26): 55-62.