The Impact of Social Media on Stock Market Empirical Analysis Based on Text Mining

Wendy Zhang

Abstract: The performance of stock market not only reflects a country's economic development and capital market trend, but also affects individual investment decision, company's business strategy, and even the national economic policy. This paper employs big data from social media to analyze and predict stock markets, utilizing text mining techniques, machine learning algorithms, and econometric methods to explore the impact of social media sentiment on stock markets. This article constructs social media sentiment index from multiple dimensions, and conduct empirical research on the performance of the stock market. The research results indicate that there exists a strong non-linear correlation between social media sentiment and stock markets, and incorporating social media sentiment indices helps explain and predict stock market behavior.

Keywords: text mining; sentiment analysis; machine learning; social media; stock market

1. Introduction

Stock markets are one of the most important components of free market economies. Stock markets provide companies with excellent financing platforms and legally guaranteed financing environments, enabling companies to efficiently raise funds needed for development and achieve large-scale operations. Investors, through purchasing stocks and stock transfers, share corporate risks while enjoying corresponding benefits such as stock appreciation and corporate dividends.

The U.S. stock market, as the most developed stock market in the world today, essentially covers all well-known enterprises globally, involving dozens of industries including energy, materials, industry, agriculture, pharmaceuticals, and consumer goods. The rich variety of stock trading types, massive number of stock issuances, and unparalleled market scale make the U.S. stock market performance not only reflect America's economic and political trends but also influence global political and economic development.

Stock markets serve as barometers reflecting economic development conditions and capital market trends. Stock market performance can reflect information about capital supply and demand conditions, industry development trends, political situation changes, and corporate operating conditions, which holds significant importance for macroeconomic policy regulation and corporate business decision-making. Therefore, analyzing and predicting stock markets can not only help investors avoid losses but also benefit national economic policy formulation and corporate business decision improvement.

Behavioral finance theory applies research findings from sociology, psychology, and behavioral studies to examine investors' irrational behavior and analyzes and explains capital market performance from investors' behavior and psychology perspectives. Behavioral finance modifies mathematical finance by incorporating investors' value perceptions, namely psychological motivations and personal behaviors, into the analysis of financial market fluctuations (Froot and Scharfstein, 1990). Psychological research indicates that emotions play important roles in people's decision-making processes. Behavioral finance, which incorporates psychological factors, believes that investors' irrational psychology and behavior affect stock market trends and cause stock prices to deviate from their intrinsic values (Kahneman and Tversky, 1997).

As carriers for people to record and express themselves, social media contains vast amounts of user emotion and behavioral information, providing abundant data sources for behavioral finance analysis. The U.S. internet penetration rate reaches as high as 88.7%, with hundreds of millions of users recording life and exchanging viewpoints on social media. Daily, Twitter publishes over 500 million posts, Instagram receives over 40 million photo uploads, and Facebook's global users exceed 2.4 billion. Simultaneously, increasing numbers of studies have noticed social media's influence, with extracting information from social media and refining investor sentiment gradually becoming hot topics in fields such as finance, computer science, and management.

Since the outbreak of the COVID-19 pandemic, stock markets in multiple countries have experienced large-scale turmoil. On March 12, the three major U.S. stock indices opened down more than 7%, collectively triggering circuit breaker mechanisms; Brazil's stock market fell 15%, triggering second-tier circuit breakers within a single day; South Korea's stock market fell nearly 11%... Currently, at least 12 countries' major stock markets have triggered circuit breaker mechanisms and suspended trading, with dozens of stock markets entering "technical bear markets." Meanwhile, public worry and anxiety emotions on Twitter and Facebook have intensified, providing sufficient reason to hypothesize correlations between social media sentiment and stock market performance.

This paper expands investor sentiment to public sentiment, based on big data from social media, utilizing web crawling, natural language sentiment analysis, machine learning algorithms, and other technologies to explore the relationship between public sentiment on social media and stock markets. This paper combines advanced computer science with cutting-edge financial theory to improve the accuracy of stock prediction models. Accurately predicting stock trends can not only help investors achieve profits but also provide reasonable bases for formulating relevant economic policies.

2. Literature Review

Stock market analysis and prediction have always been hot topics of concern in academic and business circles. Early stock market prediction mainly relied on random walk theory and efficient market hypothesis. Random walk theory believes that stock price changes are similar to the Brownian motion of random walks, with stock market fluctuations being random and irregular, making stock price movements and trends unpredictable. The efficient market hypothesis, based on rational person assumptions, similarly considers stock prices unpredictable because in efficient markets, all information has been timely, accurately, and fully reflected in stock price trends, preventing investors from obtaining excess returns above expected returns through analyzing past prices.

For stock market prediction, classical finance has proposed numerous capital pricing theories based on expected utility functions, including Markowitz's (1952) portfolio selection theory, Sharpe's capital asset pricing model, and Stephen Ross's (1990) arbitrage pricing theory. However, these stock price prediction models based on rational person assumptions cannot explain some phenomena in stock markets, such as long-term stock price reversal effects (De Bondt, 1985) and equity premium puzzles (Mehra and Prescott, 1985), with their predictions often diverging from reality.

Behavioral finance theory relaxes the rational economic person assumption, considering behavioral actors' psychological factors and thereby modifying original classical finance theories. Baker's (2007) research shows that investor sentiment can be measured, and investor sentiment fluctuations have discernible important impacts on stock markets.

The noise trading model proposed by DeLong, Shleifer, Summers, and Waldmann (1990), abbreviated as the DSSW model, is a pioneer in investor sentiment element pricing. Noise refers to information unrelated to asset value but affecting its price (BLACK F, 1986), namely the deviation between asset prices and values. Noise trading refers to transactions based on irrational psychology or information unrelated to asset value, while noise traders are irrational investors conducting noise trading. The DSSW model introduces noise traders into asset pricing models and explains the impact of noise trading on asset prices.

Commonly used investor sentiment measurement methods mainly include two steps: first, selecting market indicators such as overall market turnover rates, closed-end fund discount rates, IPO first-day premium rates, and new stock issuance numbers as sentiment variables; second, using principal component analysis, partial least squares regression methods, or dynamic factor modeling methods to construct comprehensive investor sentiment indices. Subsequently, appropriate financial theories and statistical models are selected according to different research purposes, with investor sentiment indices added as variables to complete the entire empirical analysis process.

With the widespread application of social media, many studies extract corresponding indicators from social media (Facebook, Twitter, news websites, etc.) to predict changes in various economic and business indicators. For example, D. Gruhl (2005) used content from blogs, media, and websites to predict book sales; G. Mishne (2006) conducted sentiment analysis on blog content to predict movie box office performance. Additionally, Google search indices have been proven to predict flu trends (Choi and Varian, 2009). These studies indicate that content in social media has a certain degree of predictive capability.

Behavioral finance theory believes that investor sentiment can influence stock market performance, while social media provides platforms for extensive investor communication. Therefore, many studies extract investor sentiment from social media to construct sentiment indicators. For example, Schumaker and Chen (2009) used text analysis technology to explore relationships between sentiment indices in financial news and stock comments with stock price movements; Yigitcan Karabulut (2017) used Facebook's national happiness index to predict stock price movements; Bollen J and Mao (2011) found that calm-type emotions in Twitter could predict the Dow Jones Index to some extent.

Compared to the above literature, this paper has the following two main innovations:

First, we separately use sentiment dictionary methods and five machine learning algorithms for text sentiment analysis. Results show that machine learning algorithms' accuracy rates are far higher than sentiment dictionary methods, with the Naive Bayes algorithm performing best at 88% accuracy, providing a solid data foundation for subsequent research.

Second, comparing traditional linear regression models with deep learning algorithm LSTM, we use the typical non-linear regression model LSTM (Long Short-Term Memory networks) to construct stock price prediction models with smaller errors, indicating strong non-linear relationships between social media sentiment and stock market performance.

3. Research Design

This paper comprehensively applies knowledge from finance, computer science, econometrics, and other multidisciplinary fields to explore the impact of social media on stock markets. Considering Twitter platform's monthly active user count exceeding 300 million and its broad influence, selecting Twitter as a data source to construct social media sentiment indices can relatively accurately reflect social media users' emotions. Compared to the Dow Jones Index and NASDAQ Index, the S&P 500's constituent stocks have broader coverage, stronger representativeness, and better continuity, making the S&P 500 Index a relatively precise measure of U.S. stock market performance.

In terms of theoretical analysis, this paper first clarifies the research background and significance, and through literature research, compares the current research status of related topics domestically and internationally, thereby formulating reasonable research methods and technical routes to lay theoretical foundations for subsequent empirical research. In terms of empirical research, this paper selects the globally influential Twitter platform and S&P 500 Index as research objects, utilizing text mining technology, natural language sentiment analysis, machine learning algorithms, and econometric methods to achieve data collection, processing, and model construction, thereby exploring the impact of social media sentiment on stock markets.

Technical Methodology

(1) Web Scraping Technology

This paper employs the programming language Python and web scraping technology to obtain text data published by all users on Twitter within a certain time period, serving as the data source for subsequent analysis.

(2) Natural Language Processing

This paper adopts methods such as word segmentation and bag-of-words models to clean and segment the crawled unstructured text content, and uses natural language sentiment analysis technology to determine the emotional tendencies in texts.

(3) Machine Learning Algorithms

This paper compares sentiment dictionary-based text sentiment analysis with machine learning-based text classification algorithms. The latter's accuracy rate is significantly higher than the former, so we ultimately choose the Naive Bayes algorithm to analyze the emotional tendencies of text content on Twitter.

(4) Econometric Methods

To construct social media sentiment indices and study the relationship between social media sentiment and stock markets, this paper uses numerous econometric methods for analysis, such as correlation coefficient analysis and Granger causality tests.

(5) Deep Learning Models

Traditional linear regression methods have difficulty solving multi-variable or multi-input prediction problems. LSTM-based recurrent neural networks can effectively solve multi-input variable problems, thus performing better in multivariate time series prediction. This paper uses LSTM to construct stock time series prediction models.

4. Sentiment Analysis

(I) Data Collection and Processing

1. Data Sources

According to the "Digital 2018" internet research report jointly released by internet data research institutions We Are Social and Hootsuite, over 4 billion people worldwide use the internet, with over 3 billion people using social media monthly. Social media serves as an important medium for users to express opinions and share life experiences. Content on social media can relatively authentically and accurately describe users' psychological states and behavioral characteristics. Globally renowned social media platforms mainly include Facebook, Twitter, YouTube, WhatsApp, LinkedIn, etc., while domestic social media includes Weibo, WeChat, QQ, etc.

This paper selects Twitter as the social media data source. As the world's largest social media platform, Twitter has over 300 million monthly active users. According to Internet Usage & Social Media Statistics data, Twitter users publish over 590 million tweets daily. These tweets aggregate users' opinions, attitudes, and emotions, making Twitter an excellent data source for sentiment analysis to measure social media sentiment effectively.

The association between panic and anxiety emotions about COVID-19 on social media since early 2020 and large-scale circuit breakers in stock markets worldwide demonstrates the impact of negative emotions on social media on stock markets. To increase research extensibility and exclude the influence of sudden events, selecting tweets from normal periods has more universal significance. Meanwhile, since periods that are too short would affect result accuracy, this paper ultimately selects tweets from April to September 2019, spanning six months, as research subjects.

This paper uses various data from the S&P 500 Index to measure stock market performance, as the S&P 500's constituent stocks have characteristics of broad coverage, strong representativeness, and good continuity, allowing relatively objective and comprehensive reflection of stock market dynamics. For data sources, this paper obtains S&P 500 Index volatility (VIX) from the Chicago Board Options Exchange (CBOE) and S&P 500 Index opening prices, closing prices, high and low points, and trading volumes from the famous financial website Yahoo! Finance.

This paper primarily uses the programming language Python to obtain and process data. To facilitate data analysis and processing, this paper filters out 1,716,602,742 English tweets from the crawled 2,121,821,445 tweets in various languages from around the world. This paper uses random functions to select 381,467,276 data points for subsequent analysis and processing, reducing the overall workload by 77%.

2. Text Cleaning and Processing

Text cleaning and processing are important prerequisite processes for sentiment analysis. Noise in raw data interferes with the analysis process and affects result accuracy. Since Twitter logs obtained through web crawling have complex formats and contain substantial noise such as punctuation marks, @username handles, etc., this noise information affects data analysis results, making it necessary to clean and process crawled data before sentiment analysis.

This paper first removes @username format handles and contained web links from tweets, then filters English tweets from datasets containing tweets in various languages to facilitate text analysis. After removing punctuation marks from texts, we use the Natural Language Toolkit (nltk) to split sentences into words, use its built-in WordNet dictionary to remove stopwords and meaningless words, then perform part-of-speech tagging and lemmatization on cleaned words, and finally use bag-of-words models for text feature extraction and vectorization.

The Natural Language Toolkit (nltk) is an open-source natural language processing (NLP) library written in Python, providing easy-to-use interfaces and rich text processing libraries, making it an important tool for natural language processing in Python. This paper mainly applies nltk for word segmentation, stopword removal, part-of-speech tagging, lemmatization, and other text processing tasks.

Part-of-speech tagging (POS tagging) refers to annotating each word in segmentation results with correct parts of speech, namely the process of tagging whether each word is a noun, verb, adjective, or other parts of speech. Lemmatization refers to restoring words in any form to their general forms, such as restoring "ate" to "eat." Part-of-speech tagging is the foundation of lemmatization; directly restoring words has low accuracy rates. Therefore, this paper first uses the pos_tag method in the nltk library to tag word parts of speech, then uses the WordNetLemmatizer function for restoration.

(II) Sentiment Dictionary-Based Sentiment Analysis

Sentiment dictionaries are collections of words with emotional tendencies. Sentiment dictionary-based sentiment analysis matches texts with words annotated with emotional polarities and intensities in sentiment dictionaries, then calculates comprehensively to obtain overall text emotional tendencies. The selection of sentiment dictionaries is crucial; using well-annotated sentiment dictionaries is the foundation for obtaining accurate analysis results. Currently, relatively mature and comprehensive English sentiment dictionaries mainly include GI (The General Inquirer), SentiWordNet, LIWC (Linguistic Inquiry and Word Count), etc.

This paper selects the famous sentiment analysis dictionary SentiWordNet as the dictionary source. SentiWordNet performs sentiment classification on entries in WordNet and annotates their emotional tendency weights (ESULI and SEBASTIANI, 2005). Emotional tendencies are divided into three major categories: positive, negative, and objective, with scoring for each. Since a word can have multiple parts of speech, with different meanings and uses for each part of speech, part-of-speech tagging can improve SentiWordNet recognition accuracy.

Each word's data in SentiWordNet includes part of speech, ID, positive score, negative score, synonyms, semantic labels, and synonym meanings. Since a word often contains multiple meanings, such as "good" having 4 meanings just as a noun, after part-of-speech tagging, we perform weighted statistics on word scores under the same part of speech. The formula for calculating each word's emotional score is the weighted average of n meanings, where n refers to the word's n meanings.

This paper first divides preprocessed tweets into words, performs part-of-speech tagging on words, then matches them with words in the SentiWordNet sentiment dictionary. Since a word often has multiple meanings, meanings in SentiWordNet are arranged according to common usage frequency. To objectively measure word emotions, this paper adopts two methods to calculate word scores: the first weights all meaning scores, while the second selects the first meaning's score as the word's score. Using the dictionary's built-in functions, we calculate each word's positive and negative scores, subtract them to get the word's neutral score, and sum all word scores in tweets to get that tweet's emotional score.

To evaluate sentiment dictionary classification effectiveness, this paper uses Kaggle's manually annotated sentiment tendency dataset "preprocessed-twitter-tweets" as the analysis sample. For processing convenience, this paper labels positive tweets as 1, negative tweets as -1, and neutral tweets as 0 in this dataset, converting emotional scores calculated by sentiment dictionaries to the same format for comparing sentiment dictionary classification effectiveness.

Using the first method, namely weighting all meaning scores of a word to calculate the word's score, 58% of tweets' emotional tendencies matched manual annotations. Using the second method, namely selecting the first meaning's score as the word's score and then aggregating to get tweet scores, 67% of tweets' emotional tendencies matched manual annotations. This shows that sentiment dictionary-based sentiment analysis accuracy is not high, with correctly calculated tweet proportions not even exceeding 70%.

Since sentiment dictionary-based sentiment analysis belongs to unsupervised text classification methods, directly matching words in texts with emotional words in sentiment dictionaries to calculate emotional scores, however, a word often has multiple meanings, and simple weighted calculation methods are prone to significant errors. As shown above, whether weighting meanings or directly selecting top meanings to measure words, the analysis results obtained are unsatisfactory because unsupervised classification methods cannot accurately determine word meanings in sentences. The following section will focus on discussing supervised emotional classification methods based on machine learning.

(III) Machine Learning Algorithm-Based Sentiment Analysis

1. Text Feature Representation

Computers cannot understand natural language texts, so before using machine learning algorithms to process data, we need to first convert them into numerical features that computers can recognize. Text featurization methods mainly include Set of Words models and Bag of Words models. Set of Words models use 0-1 to represent word quantities in texts, only focusing on word presence without considering specific quantities. Bag of Words models add word frequency as feature representation based on Set of Words models, making them the most commonly used method for featurizing text.

The Bag of Words model (BoW) uses machine learning algorithms for text feature representation and modeling. Feature representation refers to converting unstructured text into structured data through natural language processing and data mining technologies to facilitate subsequent model text processing. The Bag of Words model treats text as bags containing words, ignoring article word order, grammar, and syntax, not considering contextual relationships, and merely viewing text content as combinations of words or phrases, assigning corresponding weights according to word appearance frequencies.

The main processes of the Bag of Words model include word segmentation, word frequency statistics, and vectorization. After word segmentation, we count the frequency of each word appearing in texts and use this as the word's feature, then match words in texts with their frequencies to achieve vectorization, obtaining feature matrices composed of words and word frequencies. This paper applies the Bag of Words model to training sets for text feature representation and extraction, inputs them into algorithm models for training, and then evaluates and compares model classification effectiveness.

2. Training Algorithm Models

As a supervised classification method, the core step of machine learning for text sentiment analysis is using manually annotated datasets to train algorithm models. A well-annotated dataset is the foundation for accurate model training. This paper adopts Stanford University's manually annotated Sentiment140 as the training set source. This dataset contains 1.6 million tweets collected from the Twitter API, with each tweet annotated for polarity: 0 for negative, 2 for neutral, and 4 for positive.

To reasonably select training sets, this paper conducted the following experiments to explore the impact of training set quantity on model accuracy. Accuracy refers to the rate at which model predictions match actual results. I first randomly selected 10,000 positive and negative tweets each from 1.6 million tweets and used logistic regression models to train these 20,000 data points, then used the trained model to predict test sets, achieving 73.26% accuracy. Using the same method, extracting 100,000 positive and negative tweets each as training sets achieved 75.44% prediction accuracy, meaning the latter correctly predicted 218 more tweets out of every 10,000 than the former. To improve accuracy, this paper processed all 1.6 million tweets and divided them into training and test sets at an 8:2 ratio.

Machine learning-based sentiment analysis mainly includes three stages: training, testing, and application. In the training stage, this paper first separates texts and manually annotated labels, performs feature representation on texts to convert them into vectors that computers can understand, then uses algorithm models to fit extracted features and labels, obtaining classifiers based on training sets. In the testing stage, we use classifiers obtained in the training stage to predict manually annotated test sets, comparing prediction results with manual annotation results to evaluate algorithm model classification effectiveness. Finally, we select the best-performing algorithm model from the testing stage and apply it to actual datasets to predict dataset emotional tendencies.

Since classification algorithms have different theoretical foundations and implementation principles, different classification algorithms are suitable for different fields and scenarios, and classification algorithm performance varies across different datasets. Therefore, to select algorithms suitable for sentiment analysis on Twitter datasets, this paper uses Support Vector Machine (SVM), Naive Bayes Model (NB), K-Nearest Neighbors (KNN), Logistic Regression Model (LR), and Random Forest (RF) - five machine learning algorithms to fit training sets and evaluate each algorithm model's classification effectiveness, thereby selecting the most suitable algorithm model for this paper.

This paper uses the following four commonly used evaluation metrics in machine learning to assess algorithm text classification effectiveness:

Accuracy = Number of correctly predicted samples / Total number of samples
Precision = Number of samples predicted as positive / Number of samples actually positive
Recall = Number of positive samples in total samples / Number of correctly predicted samples
F1-score = 2 × Precision × Recall / (Precision + Recall)

To test the training effectiveness of each classifier, this paper first uses the Bag of Words model for text feature representation, then adopts the same 20,000 training samples and 500 test samples to train and test algorithm models respectively. The evaluation results are as follows:

Table 1: Classification Algorithm Evaluation Results Based on Bag of Words Model

Algorithm	Accuracy	Precision (neg/pos)	Recall (neg/pos)	F1 (neg/pos)
SVM	0.76	0.80/0.74	0.70/0.83	0.75/0.78
NBM	0.79	0.79/0.78	0.77/0.80	0.78/0.79
KNN	0.65	0.72/0.62	0.49/0.81	0.58/0.70
RF	0.75	0.78/0.73	0.69/0.81	0.73/0.77
LR	0.78	0.80/0.75	0.72/0.83	0.76/0.79

From the above table, we can see that the K-Nearest Neighbors algorithm performs worst, while Naive Bayes and Logistic Regression models perform best. Compared to Logistic Regression models, Naive Bayes has more balanced classification effectiveness for positive and negative sentiments, so this paper ultimately adopts the Naive Bayes model for text sentiment classification.

Table 2: Trained Naive Bayes Model Performance

Category	Accuracy	Precision	Recall	F1-score
Negative	0.88	0.88	0.88	0.88
Positive	0.88	0.89	0.88	0.89

After selecting the algorithm model, this paper uses 1.6 million annotated tweets from Sentiment140 to train algorithm models. The final sentiment classifier achieves 88% accuracy, far exceeding the 67% accuracy of sentiment dictionaries. Sentiment analysis effectiveness directly affects empirical research accuracy, so this paper ultimately adopts machine learning methods for sentiment analysis.

(IV) Constructing Social Media Sentiment Indices

After using machine learning algorithms to achieve sentiment classification of obtained texts, this paper uses Bayesian models to classify obtained tweets, thereby determining tweet emotional tendencies. After implementing sentiment classification using machine learning algorithms, we separately count positive and negative tweet quantities and use these as foundations to construct social media sentiment indices.

Since this paper's research purpose is to explore relationships between social media and stock market performance, when constructing indices, this paper draws from related research in the behavioral finance field, such as using methods by Werner Antweiler and Murray Z. Frank to construct investor sentiment indicators for building bullishness indices, and borrowing methods by Sanjiv R. Das and Mike Y. Chen to construct sentiment disagreement indices. To describe social media sentiment from multiple angles, this paper constructs the following social media sentiment indicators:

Simple Sentiment Index (SSI):
SSI = (Positive_t - Negative_t) / (Positive_t + Negative_t)

Positive Sentiment Index (PSI):
PSI = Positive_t / (Positive_t + Negative_t)

Negative Sentiment Index (NSI):
NSI = Negative_t / (Positive_t + Negative_t)

Bullishness Sentiment Index (BSI):
BSI = log(1 + Positive_t) - log(1 + Negative_t)

Sentiment Disagreement Index (DIS):
DIS = 1 - |Positive_t - Negative_t| / (Positive_t + Negative_t)

Where Positive_t represents the number of positive sentiment tweets in time period t, and Negative_t represents the number of negative sentiment tweets in time period t. After annotating sentiment through machine learning methods, this paper uses the above formulas to calculate related indices and construct social media sentiment index time series.

5. Empirical Research

(I) Correlation Analysis

Correlation analysis is a statistical analysis method for studying correlations between variables. Common correlation analysis methods include chart analysis, covariance matrix methods, correlation coefficient methods, and regression analysis methods. Since covariance is suitable for qualitative analysis (positive/negative and presence/absence of correlations) and cannot measure correlation magnitude, this paper adopts correlation coefficient methods to explore correlations between various indicators.

Commonly used correlation coefficients in statistics include Pearson correlation coefficients, Spearman correlation coefficients, and Kendall correlation coefficients. Unlike Pearson correlation coefficients, Spearman and Kendall correlation coefficients use rankings rather than specific values to reflect correlation degrees between variables. To measure correlations more precisely, this paper uses Pearson correlation coefficients for pairwise correlation testing between variables.

The Pearson correlation coefficient is a commonly used method for measuring linear correlation between variables, calculated as the covariance between variables x and y divided by the product of their respective standard deviations:

$$r = \Sigma[(xi - \bar{x})(yi - \bar{y})] / \sqrt{[\Sigma(xi - \bar{x})^2\Sigma(yi - \bar{y})^2]}$$

Pearson correlation coefficients range between -1 and 1; the larger the absolute value of the correlation coefficient, the stronger the correlation. When the correlation coefficient is 0, it indicates no association between variables. Since negative sentiment indices, sentiment disagreement indices, and positive sentiment indices are negatively linearly correlated, while bullishness indices and positive sentiment indices are positively linearly correlated, selecting positive sentiment indices can represent negative sentiment indices, sentiment disagreement indices, and bullishness indices for correlation analysis with stock market indicators. SPSS calculations of Pearson correlation coefficients between various indicators are as follows:

Table 3: Correlation Analysis

	Volume	Return	VIX	PSI	NSI	SSI	BSI
Volume	1	—	—	—	—	—	—
Return	-0.186*	1	—	—	—	—	—
VIX	0.295**	-0.299**	1	—	—	—	—
PSI	0.034	-0.001	-0.302**	1	—	—	—
NSI	-0.034	0.001	0.302**	-1.000**	1	—	—
SSI	0.037	-0.015	-0.283**	0.867**	-0.867**	1	—
BSI	0.034	-0.001	-0.302**	1.000**	-1.000**	0.868**	1
DIS	-0.034	0.001	0.302**	-1.000**	1.000**	-0.867**	-1.000**

**p < 0.01 (two-tailed), *p < 0.05 (two-tailed)

From correlation analysis results, we can see that Positive Sentiment Index (PSI), Simple Sentiment Index (SSI), and Sentiment Disagreement Index (DIS) are significantly negatively correlated with S&P 500 Index volatility (VIX). When positive sentiment indices rise, S&P 500 Index volatility decreases; when sentiment disagreement indices rise (i.e., when negative or positive emotions on social media are relatively dispersed), S&P 500 Index volatility increases, which aligns with facts.

(II) VAR Model and Granger Causality Test

The VAR model (Vector Autoregression Model) is an econometric model used to estimate dynamic relationships among jointly endogenous variables, commonly used for analyzing and predicting multiple related economic indicators. The VAR model extends standard autoregression (AR) models from single time series to multivariate time series. In VAR models, variables in the same sample period are regressed against lagged terms of all variables to construct models. A p-order VAR model can be expressed as:

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + ... + A_p Y_{t-p} + \varepsilon_t$$

Where A_k $(k=1,2,...,p)$ represents coefficient matrices of vector autoregression, and ε_t can be viewed as Gaussian noise.

The prerequisite for constructing VAR models is stationary time series. Common methods for stationarity testing include DF tests, ADF tests, and PP tests. This paper uses ADF tests for stationarity testing of related indicators. Test results and standards are as follows:

Table 4: ADF Test Critical Values

Confidence Level	Critical Value
1%	-3.513
5%	-2.892
10%	-2.581

Table 5: S&P 500 Index ADF Test

Indicator	t	p-value	Stationary
Opening Price	-1.775	0.3928	No
Closing Price	-2.524	0.1096	No
Volume	-4.789	0.0001	Yes
Volatility	-3.375	0.0106	Yes
Return	-11.722	0.0000	Yes

Table 6: Social Media Sentiment Index ADF Test

Indicator	t	p-value	Stationary
Positive Sentiment Index (PSI)	-6.815	0.0000	Yes
Negative Sentiment Index (NSI)	-6.815	0.0000	Yes
Simple Sentiment Index (SSI)	-7.414	0.0000	Yes
Bullishness Index (BSI)	-6.838	0.0000	Yes

Sentiment Disagreement Index (DIS)	-6.815	0.0000	Yes

From the above tables, we can see that for S&P 500 Index ADF tests, opening and closing price time series are not stationary sequences; volume and return time series are stationary at 1% confidence levels; volatility time series is stationary at 5% confidence levels. For social media sentiment index ADF tests, PSI, NSI, SSI, BSI, and DIS time series are all stationary at 1% confidence levels.

This paper constructs VAR models for stationary time series Volume, Return, VIX, PSI, BSI, and SSI after z-score standardization, and uses Granger causality tests and impulse response functions to evaluate and analyze VAR results. The Granger Causality Test is a commonly used method for analyzing Granger causal relationships between economic variables, with its statistical essence being prediction of stationary time series.

This paper first determines optimal lag orders for Granger causality tests according to minimum AIC and SC value criteria. If AIC and SC do not simultaneously reach minimum values, we select orders according to minimum LR test results. Generally, 4 orders are used as maximum lag orders; causal relationships detected with lag orders exceeding 4 orders often lack significance.

The null hypothesis is: X does not Granger-cause Y. At 95% confidence levels, Granger causality tests on variables yield the following results:

Table 7: Granger Causality Test Results

Y	X	Chi2	df	Prob>chi2
VIX	PSI	16.888	4	0.002
	BSI	16.469	4	0.000
	SSI	63.312	4	0.000
PSI	VIX	90.092	4	0.000
	Volume	53.619	4	0.000
	Return	72.923	4	0.000
Volume	PSI	103.930	4	0.000
	BSI	103.690	4	0.000
	SSI	66.770	4	0.000
BSI	VIX	88.461	4	0.000
	Volume	52.760	4	0.000
	Return	480.090	4	0.000
Return	PSI	47.718	4	0.000
	BSI	47.445	4	0.000
	SSI	31.649	4	0.000
SSI	VIX	50.604	4	0.000
	Volume	42.235	4	0.000
	Return	41.999	4	0.000

Since NSI and DIS are linearly correlated with PSI, Granger causality test results are identical to PSI, with p-values of 0.000. Combining table data, we find that all variable Granger causality test p-values are less than 0.05, indicating that social media sentiment indices and S&P 500 Index volatility, volume, and returns are mutually Granger causal at 95% confidence levels. This shows that adding social media sentiment indices helps predict S&P 500 Index changes, and S&P 500 Index changes also help explain social media sentiment index variations.

(III) Impulse Response Analysis

Granger causality tests describe causal relationships between variables but can only indicate whether one variable helps explain and predict other variables, without determining action directions and influence timing between variables. Impulse response analysis applies "exogenous shocks" to variables in VAR models, observes dynamic impacts on other variables in models, thereby determining change trends between variables.

This paper first fits variables using VAR models, then uses impulse response functions to predict 8-period changes. Impulse response functions measure impacts of one standard deviation shocks in random disturbance terms on endogenous variables. Since dependent variables in impulse response functions are all affected by shocks from variables sequentially before them, without being influenced by variables sequentially after them, to avoid such interference, this paper minimizes variables in functions. Based on variable dynamic changes over time periods after receiving shocks, we draw impulse response diagrams. The following are two illustrative diagrams:

Figure 4: NSI-VIX Impulse Response Diagram

Each row in impulse response diagrams shows impacts of different variables receiving shocks from the same variable, while each column shows impacts of the same variable receiving shocks from different variables. The horizontal axis represents forecast periods; in the above diagram, 8 periods represent impacts brought by shocks over the next 8 periods.

From the NSI-VIX impulse response diagram, we can see that impacts between VIX and NSI concentrate in the first two periods, then gradually converge. Volatility shocks cause significant increases in negative sentiment indices, which aligns with actual situations.

This paper conducts pairwise impulse response diagram analysis for three S&P 500 Index variables and five social media sentiment variables, ultimately determining mutual influence relationships between variables. As shown in the two illustrative diagrams above, influence relationships between variables mostly concentrate in the first two periods, meaning impacts brought by shocks are often temporary and do not last long. This also aligns with the strong timeliness characteristics of social media and S&P 500 Index.

(IV) Multivariate Linear Regression Model Prediction

To study the explanatory and predictive capabilities of social media sentiment indices for the S&P 500 Index, this paper uses Python to conduct regression analysis on social media sentiment indices and S&P 500 Index. As known from above, among S&P 500 Index indicators, returns, volume, and volatility are stationary sequences, while among social media sentiment indices, PSI, BSI, and SSI are stationary sequences without linear relationships. Therefore, we respectively use returns, volume, and volatility as independent variables and other indicators as dependent variables for regression.

This paper first divides all data into training and test sets, uses linear regression models to fit training sets, then uses trained linear regression models to predict test sets. Linear regression models assume linear correlations between variables, assign regression coefficients according to variable weights, then calculate linear regression equations combining intercepts and regression coefficients.

Taking return prediction as an example, we first conduct regression analysis without adding social media sentiment indices. Through linear regression models, we calculate an intercept of 0.027 and variable coefficients as open: -2.503, high: -0.060, low: 0.327, close: 2.264. The linear regression equation is:

Return = 0.027 - 2.503×Open - 0.060×High + 0.327×Low + 2.264×Close

Comparisons between predicted and actual values are as follows:

Figure 5: Multivariate Linear Regression Predicted Returns vs. Actual Returns

This paper uses Root Mean Squared Error (RMSE) to evaluate regression results. This regression equation's RMSE is 0.4435. After adding PSI and NSI, the regression equation's RMSE becomes 0.4302. Smaller RMSE indicates higher fitting degrees, so social media sentiment indices help explain and predict the S&P 500 Index.

(V) LSTM Neural Network-Based Time Series Prediction

LSTM Long Short-Term Memory networks are types of temporal recurrent neural networks and variants of RNN (Recurrent Neural Networks). LSTM's design structure aims to solve long-term dependency problems existing in general RNNs, making LSTM more suitable than RNN for processing and predicting time series with long intervals and delays.

Traditional linear regression methods have difficulty solving multi-variable or multi-input prediction problems. LSTM-based recurrent neural networks can effectively solve multi-input variable problems, thus performing better in multivariate time series prediction. As typical non-linear models, LSTM can also serve as complex non-linear units to construct larger deep neural networks.

Time series prediction analysis refers to using past variable values over periods to predict future values, representing relatively complex prediction modeling problems. Unlike simple regression analysis, time series prediction depends on the sequential order of variable values; changing variable value sequences completely alters prediction results.

Since multivariate linear regression model prediction results are unsatisfactory, this paper adopts non-linear LSTM models for stock data prediction. To facilitate comparison with multivariate linear regression model prediction results, this paper uses identical datasets and adopts the same standardization methods for data processing.

This paper uses LSTM model hidden layers with 50 neurons for deep learning on data. After completing training and testing, we use MAE (Mean Absolute Error) as loss functions to draw loss curves. Loss functions describe gaps between predicted and true values; smaller loss values indicate predicted values closer to true values.

Figure 6: Training Set and Test Set Loss Curves

As shown above, with increasing neuron numbers, training and test set losses first decrease then converge. This occurs because models are still learning early on, causing losses to gradually decline. Later, training and test losses gradually converge and approach 0, indicating models have stopped training. This paper's test losses are smaller than training losses, suggesting LSTM models may overfit training sets.

Using the same measurement method as multivariate linear regression models, we adopt Root Mean Squared Error (RMSE) to evaluate LSTM model prediction results. When independent variables are opening prices, closing prices, high points, and low points, with dependent variables being returns, RMSE is 0.154, far lower than multivariate linear regression model RMSE. Adding PSI, NSI, and BSI to independent variables reduces RMSE to 0.120, indicating that social media sentiment helps predict return time series.

6. Research Conclusions and Implications

Asset pricing has always been a core problem in finance. Classical finance constructs a series of asset pricing models under theoretical frameworks of equilibrium pricing and arbitrage-free pricing. Behavioral finance incorporates behavioral actors' psychological factors based on classical finance, modifying and perfecting original models. As carriers for people to express opinions and share life experiences, social media contains enormous amounts of user psychological and behavioral characteristics. Combining big data from social media with behavioral finance theory to analyze stock market trends not only fully utilizes social media's information-rich advantages but also effectively explores public sentiment's impact on stock markets.

This paper selects Twitter texts and S&P 500 Index from April 1 to September 30, 2019, as data sources. Based on behavioral finance theory and applying machine learning methods and econometric models for data analysis and processing, we reach the following conclusions:

First, this paper separately uses sentiment dictionary methods and multiple machine learning algorithms for text classification and sentiment annotation of obtained Twitter texts, using identical test sets to evaluate sentiment analysis results. Machine learning-based sentiment analysis achieves 75% average accuracy, far higher than sentiment dictionary-based sentiment analysis accuracy of 67%. Among the five tested machine learning algorithms, Naive Bayes models perform best, with trained Naive Bayes models achieving 88% accuracy. Therefore, this paper adopts Naive Bayes models for text classification.

Second, this paper draws from related research in behavioral finance to construct social media sentiment indices, adopting multiple empirical research methods to explore relationships between social media sentiment indices and stock markets. In correlation analysis, this paper uses Pearson correlation coefficients to measure linear relationships between variables, finding significant correlations between social media sentiment indices and S&P 500 Index volatility. In dynamic relationship analysis, this paper employs multiple econometric methods including VAR models, Granger causality tests, and impulse response analysis. Results show that social media sentiment indices help explain and predict S&P 500 Index changes, but only short-term influence relationships exist between them.

Third, this paper separately uses multivariate linear regression models and LSTM neural networks for time series prediction. Both prediction results indicate that models incorporating social media sentiment indices have smaller errors and higher fitting degrees, so social media sentiment indices help explain and predict stock market performance. Compared to multivariate linear regression models, LSTM neural network-based predictions perform better, suggesting possible strong non-linear relationships between variables.

These research findings demonstrate that social media sentiment analysis can significantly enhance stock market prediction accuracy. The integration of advanced computational techniques with financial theory provides valuable insights for investors, policymakers, and financial institutions, improving market understanding and forecasting capabilities. The strong non-linear relationships identified between social media sentiment and stock market performance highlight the importance of using sophisticated analytical tools beyond traditional linear models.

References

[1] Froot, K. A., Scharfstein, D. S., & Stein, J. C. (1990). Herd on the Street: Informational Inefficiencies in A Market with Short-Term Speculation. National Bureau of Economic Research Cambridge, Mass. USA.

[2] Data source: Internet Live Stats - Internet Usage & Social Media Statistics.

[3] Kahneman, D., & Tversky, A. (1997). Prospect theory: An analysis of decision under risk. Econometrica: Journal of the Econometric Society, 263-291.

[4] Osborn, H., Engineers S.o.A. (1964). Latest Developments in High Frequency Welding. Society of Automotive Engineers.

[5] Markowitz, H. M. (1952). Portfolio selection. Journal of Finance, 7(1), 77-91.

[6] Ross, J., Eady, E., Cove, J., et al. (1990). Inducible erythromycin resistance in staphylococci is encoded by a member of the ATP-binding transport super-gene family. Molecular Microbiology, 4(7), 1207-1214.

[7] De Bondt, W. F. M. (1985). Does the stock market overreact to new information? Journal of Finance, 40(3), 793-805.

[8] Mehra, R., & Prescott, E. C. (1985). The Equity Premium: A Puzzle. Journal of Monetary Economics, 15(2), 145-161.

[9] Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of Economic Perspectives, 21(2), 129-151.

[10] DeLong, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. Journal of Political Economy, 98, 703-738.

[11] Black, F. (1986). Noise. Journal of Finance, 41(3), 528-543.

[12] Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. In KDD '05: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (pp. 78-87). ACM Press.

[13] Mishne, G., & de Rijke, M. (2006). Capturing global mood levels using blog posts. In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (pp. 145-152). The AAAI Press.

[14] Choi, H., & Varian, H. (2009). Predicting the Present with Google Trends. Technical report, Google.

[15] Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. ACM Transactions on Information Systems, 27(2), 1-19.

[16] Karabulut, Y. (2017). Can Facebook Predict Stock Market Activity? SSRN Electronic Journal. DOI: 10.2139/ssrn.2017099.

[17] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

[18] Banerjee, A. V. (1992). A Simple Model of Herd Behavior. Quarterly Journal of Economics, 107(3), 797-817.

[19] Sullivan, D. (2000). The Need for Text Mining in Business Intelligence. DM Review.

[20] MacDonald, C., & Ounis, I. (2006). The TREC Blogs 06 collection: Creating and analysing a blog test collection. University of Glasgow, Department of Computer Science.

[21] Seki, Y., Evans, D. K., Ku, L. W., et al. (2007). Overview of opinion analysis pilot task at NTCIR-6. In Proceedings of the Workshop Meeting of the National Institute of Informatics Test Collection for Information Retrieval Systems (pp. 265-278). National Center of Science.

[22] Salton, G., Wong, A., & Yang, C. S. (1995). A Vector Space Model for Automatic Indexing. Communication of the ACM, 18, 613-620.

[23] Quinlan, J. R. (1993). Constructing Decision Tree in C4.5. Morgan Kaufman Publishers, Programs for Machine Learning, 17-26.

[24] Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In Proc. of the 14th Intl. Conf. on Machine Learning ICML 97 (pp. 412-420).

[25] Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.

[26] McCallum, A., & Nigam, K. (1998). A Comparison of Event Models For Naïve Bayes Text Classification. Just Research.

[27] Mladenic, D. (1998). Machine Learning On Non-Homogeneous, Distributed Text Data. Doctoral Dissertation, University of Ljubljana.

[28] Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd International Conference on Knowledge Capture (pp. 70-77). ACM Press.

[29] Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In Proceedings of the National Conference on Artificial Intelligence (pp. 755-760). AAAI Press.

[30] Stone, P. (1968). The general inquirer: A computer approach to content analysis. Journal of Regional Science, 8(1), 113-116.

[31] Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss analysis. In Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (pp. 617-624). ACM.

[32] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up: sentiment classification using

machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (pp. 79-86).

[33] Wilson, T., Hoffmann, P., Somasundaran, S., et al. (2005). OpinionFinder: A system for subjectivity analysis. In Proceedings of HLT/EMNLP on Interactive Demonstrations (pp. 34-35). Association for Computational Linguistics.

[34] Gamon, M., Aue, A., Corston-Oliver, S., et al. (2005). Pulse: Mining customer opinions from free text. In Proceedings of the 6th International Symposium on Intelligent Data Analysis (pp. 121-132). Springer-Verlag.

[35] Hart, R. P. (2000). DICTION 5.0: The Text-analysis Program. Sage.

[36] Vapnik, V. (1998). Statistical learning theory (Vol. 3). Wiley.

[37] Hsieh, W. W. (2009). Machine learning methods in the environmental sciences: Neural networks and kernels. Cambridge University Press.

[38] Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. The Journal of Finance, 59(3), 1259-1294.

[39] Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science, 53(9), 1375-1388.