Use Investor Sentiment to Predict Future Price Movement in the U.S. Stock Market Based on the Financial Social Media Platform

Yulun Han

May 30, 2022

1 Introduction

With the rapid development of Internet technology, social media, mainly microblogs, social news, and online social platforms, are flooding the entire Internet space at a rapid pace. The internet has evolved from a simple technical platform for information dissemination to a major carrier of social media, developing into a social network for interactive information dissemination, sharing, communication, and collaboration. The geometric growth of information volume of online public opinion, the information dissemination mechanism different from traditional media, and the interpersonal network are intertwined to bring the direct effect of the rapid expansion of influence. The influence of online information on the stock market is also becoming more and more prominent. More and more listed companies have opened official microblogs, sometimes releasing news through microblogs. While investors also rely more and more on network information to make investment decisions, they exchange and discuss on online financial platforms, and also post information on microblogging platforms or comment on others' information, forming interactive communication with listed companies and other investors. However, in traditional models of stock price prediction, researchers have often neglected to study investor sentiment. As a result, in this research, I explore the following two questions in detail. The first question is whether stock sectors can be classified according to the level of investor interest. Second, the results of the classification of stocks based on social network analysis are used to predict the future prices of stocks using the investor sentiment score.

2 Data

For the data of this research, there are two parts to the dataset, which are social media data and market data. The research focuses on 50 stocks that are the holdings of Invesco S&P 500 Top 50 ETF (ticker: XLG). The S&P 500 Top 50 consists of 50 of the largest companies from the S&P 500, reflecting U.S. mega-cap performance. The

50 target stocks belong to five different sectors, each sector containing 10 stocks. The five sectors are communication services, consumer discretionary, financials, health care, and information technology. In addition, there are two main sources of research data, which are Yahoo finance and StockTwits.

2.1 Social Media Data

Social media data is collected from 50 target companies' posts in the most recent 50 trading days from StockTwits which is a social media platform designed for sharing ideas between investors, traders, and entrepreneurs. The dataset of social media data includes user id, the text of users' posts, and post date. The average data size of one stock is about 10,160 posts in fifty trading days. The size of the entire social media dataset is approximately 508,000 posts. Among them, Tesla has the highest number of posts, with nearly 170,000 posts in fifty trading days, and Fox has only 120 posts.

The method of data collection for social media data is to use a web crawler to obtain information. 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.

2.2 Market Data

The market data consists of historical daily data from five target sector index, which are S&P 500 Consumer Discretionary index (ticker: ^SP500-25), S&P 500 Health Care index (ticker: ^SP500-35), S&P 500 Financials index (ticker: ^SP500-40), S&P 500 Info Tech index (ticker: ^SP500-45), and S&P 500 Communication Services (ticker: ^SP500-50). S&P Sector and Industry Indices measure segments of the U.S. stock market as defined by GICS. This part of the data is downloaded from Yahoo Finance.

3 Methodology and Results

For this research, I involve four computational methods, which are graph representation, community detection, sentiment analysis, and random forest regression model. According to graph representation, I create a social network graph to represent the relationship between 50 target stocks. Based on the result of graph representation, a new sector classification of stocks is derived from the analysis of

community detection. In addition to this, I built a machine learning model for predicting stock prices by using investor sentiment scores, trained on data from each of the five stock sectors respectively. In the prediction model, I use sentiment score as the independent variable. The sentiment score is calculated by performing sentiment analysis on the posts.

3.1 Graph Representation

3.1.1 Method

A social network is a collection of social actors and the relationships they ask about. A social network consists of many nodes and edges that connect these nodes. The nodes can be any unit of social analysis, such as individuals, groups, organizations, and communities. Edges represent relationship between nodes. Therefore, social network analysis is a set of theories, methods, and techniques that include the whole process of measuring and investigating the characteristics of each part of a social system and the relationships between them, representing them in the form of a network, and then analyzing the patterns and characteristics of their relationships. In this research, I defined 50 target stocks as nodes and the number of users who jointly commented as edges. When every two stocks have 23 or more co-users, an edge will be generated between the two stocks. The relationship between stocks is shown in Figure 1.

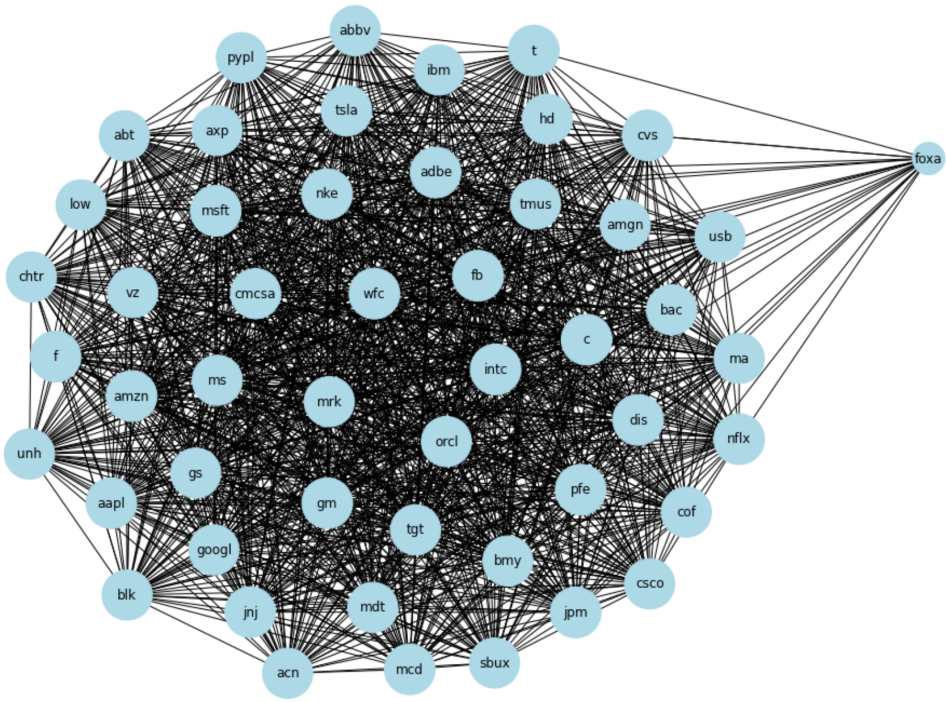

Figure 1 Graph Representation for 50 Target Stocks

3.1.2 Results

Using the graph representation method, I came up with the presence of 50 nodes and

1196 edges in the graph. Among them, it is obvious that 49 nodes are tightly connected, except for "foxa". However, "foxa" is the only one that does not appear in the cluster. The reason for this result may be that “foxa” has the lowest number of posts compared to the other 49 stocks, which means that only a small number of users participate in commenting on Fox Corporation’s stock, thus causing “foxa” to appear distant in the graph. The result obtained based on the graph representation lay the foundation for subsequent community detection to classify the stock sectors.

3.2 Community Detection

3.2.1 Method

Community detection is an important research content in the field of social network analysis and has received extensive attention from researchers in various fields in recent years. Communities are dense groups in networks that satisfy relatively tight links between nodes within the same community and relatively sparse links between nodes in different communities. Due to this characteristic of community structure, community identification is widely used in application systems such as sentiment analysis and commodity recommendation.

Based on the results obtained from the graph representation, I decided to reclassify the 50 target stocks using the community detection with the method of the second eigenvector of the Laplacian matrix. The basic principle of the Laplacian matrix is that, Laplacian Matrix = Degree Matrix - Adjacency Matrix. Based on the result of the second eigenvector, If the second eigenvector is negative or 0, the element will be a community; when the second eigenvector is positive, the element will be another community. Thus, using the above algorithm, I arrive at a new classification of stock sectors.

3.2.3 Results

As can be seen from the Figure 2, the target stocks are reclassified into five sectors and the ten stocks in each sector are closely related. In the new sector classification of stocks, the first sector includes “axp”, “cmcsa”, “cvs”, “dis”, “googl', “gs”, “ibm", “nflx”, “tgt”, “vz”. In the first community, the major sector of stocks is in communication services, so I define the first community as communication services sector. The second community includes “amzn”, “f”, “fb”, “foxa”, “intc”, “jnj”, “mcd”, “msft”, “t”, “tsla”. In the second community, the majority of stocks belongs to consumer discretionary sector, so I define the second community as consumer discretionary sector. According to the method of taking the majority, I defined the remaining three communities. The financials sector which is the third community includes “abt”, “blk”, “cof”, “csco”, “hd”, “low”, “ma”, “ms”, “tmus”, “wfc”. The health care sector which is the fourth community includes “aapl”, “abbv", “amgn”,

“chtr”, “gm”, “mrk”, “orcl”, “sbux”, “unh”, “usb”. The information technology sector which is the fifth community includes “acn”, “adbe”, “bac”, “bmy”, “c”, “jpm”, “mdt", “nke”, “pfe”, “pypl”.

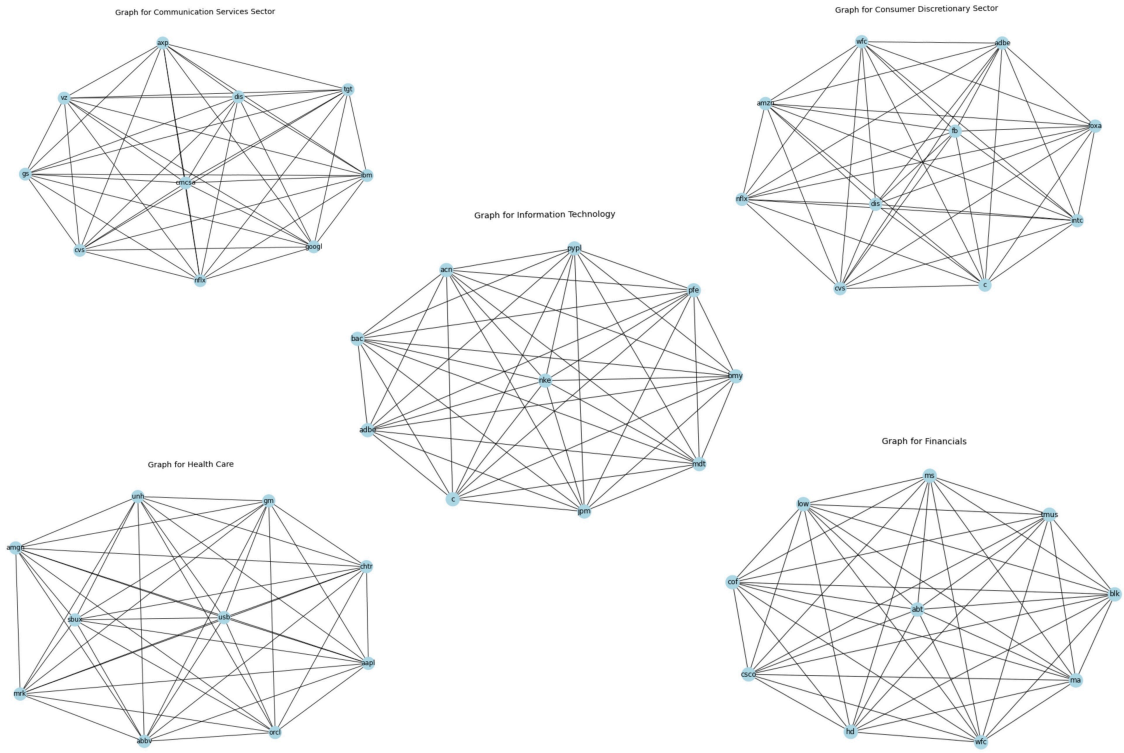

Figure 2 Sub-Graphs after Community Detection

3.2.2 Evaluation

By calculating the average clustering coefficient, average node degree, and graph density for each of the five communities, I found that each community has an average clustering coefficient of 1 and a graph density of 1. Such results imply that the nodes in each community are all fully connected, proving that the results of community detection hold. The results of the community detection also indicate that investors tend to follow or invest in stocks in the same sectors, and investor sentiment can be used in the stock price prediction model.

Evaluation Result for Community Detection

	Average Clustering Coefficient	Average Node Degree	Graph Density
Communication Services	1.000000	9.000000	1.000000
Consumer Discretionary	1.000000	9.000000	1.000000
Financials	1.000000	9.000000	1.000000
Health Care	1.000000	9.000000	1.000000
Information Technology	1.000000	9.000000	1.000000

Table 1 Evaluation Results for Community Detection

3.3 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 fifty trading days. Text mining includes a series of contents, including sentiment analysis, text classification, topic model, etc. This research uses sentiment analysis to compute average polarity as 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, and then delete duplicate posts. 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 polarity for each post by using the following formulas:

$$Polarity = \frac{Freq_{pos} - Freq_{neg}}{Freq_{pos} + Freq_{neg}}$$

After obtaining the polarity of each post, I calculated the average polarity of each day's post for each stock to represent its daily sentiment score. In the next stock price prediction model, the sentiment score will appear as an independent variable in the model.

3.4 Random Forest Regression

3.4.1 Method

To better predict the stock price and solve the problem of low prediction accuracy, this research builds a random forest regression model to predict the stock price based on investor sentiment. The stock prediction model is constructed for each of the five sectors, in which the daily sentiment score is the independent variable, and the daily adjusted closing price of the stock sector index is the dependent variable. Both variables focus on the recent 50 trading days. In addition, a randomized grid search cross-validation, a hyper tuning process is used to validate the building, fitting, and training of the prediction model. After prediction, error analysis is used to determine the performance of the model and the accuracy of the predicted values.

3.4.2 Evaluation

The root means square error (RMSE) is the standard deviation of the residuals (prediction errors). The residuals measure the distance of the data points from the regression line. the RMSE is a measure of how these residuals are distributed. In other words, it tells how the data are concentrated on the most appropriate straight line. It is also the square root of the MSE. the lower the RMSE value, the better the performance.

The mean squared error (MSE) is taken as the sum of the absolute values of the errors. The mean squared error also determines the performance of the model. In this case, errors larger than the MAE can be well noted. the lower the MSE value, the higher the prediction accuracy.

The mean absolute error (MAE) measures the average size of the errors in a set of predictions, regardless of their direction. It is the average absolute difference between predicted and actual observations, where all individual differences have equal weight. Most notably, it measures the distance between the actual and predicted values. However, the MAE does not penalize errors in prediction. So, if the error is to be considered, it should be the mean squared error or root mean squared error. The lower the value, the better.

Model Performance for Communication Services Sector:

	MAE	MSE	RMSE	Accuracy (%)
Yahoo Finance	13.918	227.698	15.090	93.840
Community Detection	9.429	145.388	12.058	95.960

Model Performance for Consumer Discretionary Sector:

	MAE	MSE	RMSE	Accuracy (%)
Yahoo Finance	67.294	5956.711	77.180	95.180
Community Detection	56.308	4012.148	63.342	96.080

Model Performance for Finanicals Sector:

	MAE	MSE	RMSE	Accuracy (%)
Yahoo Finance	16.660	400.613	20.015	97.240
Community Detection	16.760	421.725	20.536	97.350

Model Performance for Health Care Sector:

	MAE	MSE	RMSE	Accuracy (%)
Yahoo Finance	38.140	2,414.086	49.133	97.630
Community Detection	27.558	1,016.714	31.886	98.280

Model Performance for Information Technology Sector:

	MAE	MSE	RMSE	Accuracy (%)
Yahoo Finance	75.922	9395.083	96.928	97.170
Community Detection	90.703	11149.413	105.591	96.660

Table 2 Model Performance for 5 Sectors

3.4.3 Results

As the Table 2 shows, the stock price prediction model of the health care sector has the highest accuracy, and the communication services sector has the lowest accuracy.

Only in the predictive model for the information technology sector, the accuracy of the predictive model based on the sector classification provided by Yahoo Finance is higher than that of the sector classification by the community detection, and in the remaining four sectors, the prediction model based on the community-detected sector classification has higher accuracy. Although the results of the error analysis for all models are relatively large, the results of the accuracy of the model predictions remain meaningful because of the limited nature of the data.

4 Conclusion and Future Work

4.1 Conclusion

To review the research questions of this study, first I re-classified 50 stocks by sector based on investor postings through a community detection algorithm to its to demonstrate whether investors tend to be following or investing in stocks in the same sector. The conclusion shows that when every two stocks have 23 and more investors commenting, the fifty target stocks can be classified into five categories by Laplacian matrix analysis. The stock prices are predicted by the random forest regression model according to the sector classification of Yahoo Finance and the sector classification of community detection, respectively. It is found that the accuracy of the prediction model based on the community-detected sector classification is generally greater than the accuracy of the industry classification provided by Yahoo Finance. Thus, this research demonstrates that investor sentiment can reasonably predict stock prices.

4.2 Future Research

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 a frequency dictionary is too rough to get a more accurate value. Because the posts' content on social media is short and incomplete and requires 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 research, only 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 flaw in the research. The data cycle we used is relatively short, which is a fluctuation period within the recent 50 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 the United States' stock market, I believe that the data analysis based on nearly 140 consecutive trading days in 6 months is relatively reliable.