Proceedings of ICEMGD 2024 Workshop: Decoupling Corporate Finance Implications of Firm Climate Action
DOI: 10.54254/2754-1169/102/2024ED0081

Application of Sentiment Analysis to Explore the Connection Between Public Sentiment and Stock Price Movement by Python

TszYan Lau[1,a,*]

[1]*University of Toronto, Toronto, Canada*
a. tszyanleon.lau@mail.utoronto.ca
**corresponding author*

Abstract: This study aims to research the relationship between sentiments of public and the trend of the stock. Specially, it focuses on analyzing how positive or negative sentiment correlated to the trend of stocks. In methodology, this research investigates the predictive power of public sentiment on stock market trend by analyzing user's comments on Reddit by utilizing Python Reddit API Wrapper (PRAW) to systematically web-scrape relevant financial discussion from a subreddit called 'stocks'. After the data collecting, preprocessing step is conducted to removing noise and standardize text. Eventually, sentiment analysis is applied to analyze the sentiments of comments by using Natural Language Processing (NLP) techniques. This approach allows for an in-depth examination of the correlation between the sentiments expressed in Reddit comments and subsequent stock price fluctuations by comparing the outcome and the actual price movement. Eventually, through the comparison between the sentiment analysis result and actual stocks movement, the sentiment of public comments shares the same trend with the market movement.

Keywords: Sentiment Analysis, Python, NLTK, NLP

1. Introduction

A stock represents the ownership of a corporation or organization, which entitles the holder to a claim on part of the assets and earnings of the company. Therefore, aside from the issuance of bonds, releasing a specific number of shares on the stock market through investment banking is also a crucial financial tool for most corporations. Correspondingly, investors often seek to hold stocks of companies with growth prospects since this allows them to profit through the strategy of buying low and selling high. Thus, predicting the movement of stocks is always the most important step in the investment process. However, predicting stock trends correctly and accurately is always a significant challenge. In the past, due to the lack of advanced computer equipment and a significant information gap, individuals mostly relied on word of mouth or newspapers to learn about a company's development and decide whether to hold its stocks. This situation changed with the advent of the information revolution in the late 20th century. With the development of technologies like computers and data storage, more individuals prefer to collect, utilize, and analyze data to predict stock trends, using tools such as Candlestick Charts, Rate of Change, comparison of Moving Average lines, and the Stochastic Indicator. However, despite many years of growth, the one constant is that humans

DOI: 10.54254/2754-1169/102/2024ED0081

always dominate decision-making in all investing activities. Nevertheless, as is well known, humans are creatures who are easily swayed by their emotions when making decisions, meaning that most find it difficult to remain rational. According to research, happy participant would believe that the positive events have a higher percentage to occur, and the occurance of negative events will be more lesser. In contrast, sad participants would hold that negative events were more likely than positive events [1]. In the same vein, participants who were in a positive emotional state displayed optimism by overestimating their chances of winning compared to losing [2]. Therefore, when investors are optimistic to a stock, they are likely to overestimate its value and choose to hold shares of that company. However, as more and more people share the same investment decisions and thoughts, this can ultimately lead to increase in the stock price. Thus, collecting and understanding public opinion may also be a method to anticipate stock market trends. Given the crucial role of emotion in investment decisions, there is a pressing need to understand how collective sentiment expressed online influences stock prices. Sentiment analysis provides a potential and great solution by enabling the systematic examination of public emotions and opinions shared on digital platforms. Through fully utilizing the function of Natural Language Processing (NLP) technology, it can provide insights into general market sentiment and help to predict market trends more accurately and effectively. Eventually, after the outcome is generated, the comparison between the outcome and actual market movement is conducted and find the connestion of them at the end.

2. Model Formulation

This research mainly targets to find out the relationship between publish sentiment and trend of the stock price through using Python. Specially, it analyze how positive or negative sentiment correlated to the trend of stocks. This will be achieved by collecting data from Reddit, followed by preprocessing this data, and finally conducting sentiment analysis using Natural Language Processing (NLP).

2.1. Choose of Data Resource - Reddit

Reddit is one of the largest American social discussion platforms. The registered users can share the post include videos, texts, links, survey, and some interesting links, and the users can vote ‘like’ or ‘dislike’ on the post so that it will go up or down on the page, which depends on how heat the post is. Then, based on the subject of the posts, they are organized and appear in different user-created boards called ‘communities’ or ‘sub-reddit’, which can cover many different topics such as games, sports, music, daily news, financial market news [3]. According to Figure 1 which is company-published data in Q4 2023, Reddit has 73.1 million unique users all over the world [4], and the trend is still increasing. According to the statistics, from Q4 2022 to Q4 2023, the daily active user has increased 27.13% [5]. This growth underscores Reddit's role as a valuable source of data, and user comments that posted on Reddit can be considered as the reflection of public opinion and sentiment.

DOI: 10.54254/2754-1169/102/2024ED0081

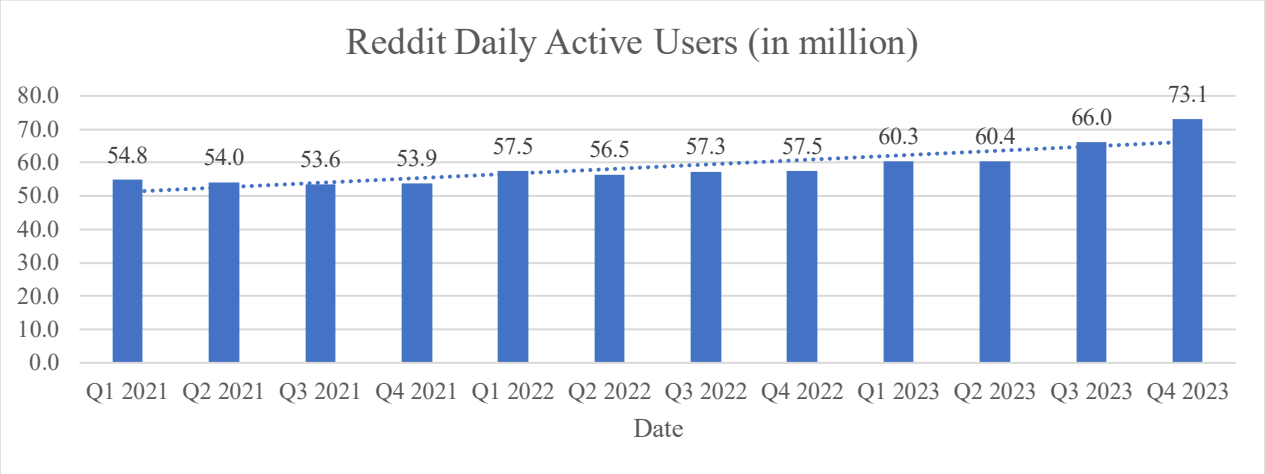

Figure 1: Reddit Daily Active Users (in million)

2.2. Data Scrapping - PRAW

The term "PRAW" refers to "The Python Reddit API Wrapper", which is a Python package that can be used to simply access Reddit API . Through utilizing PRAW in python, it helps to effectively and efficiently data mining and web scraping the comments on Reddit in a licensed way. Therefore, in this study, PRAW will be installed and used for collecting the data from the users' post. Meanwhile, financial related sub-reddit and post also should be chosen for the further text and sentiment analysis to find out the connection between sounds from the public and the alternation of price of a stock. Consequently, this paper aims to analyze comments from the post titled 'Reddit pops 38% in NYSE debut to open at $47 per share' in the 'stocks' subreddit. The target is to assess public opinion following Reddit's IPO and correlate these sentiments with changes in stock price.

2.3. Preprocessing

The preprocessing step is extremely crucial in the whole sentimental analysis process because it not only removes the 'noise' in the dataset but also enhances the accuracy of the result. At the meantime, the Python package 'NLTK' has applied in the preprocessing step. 'NLTK' is a short term of Natural Language Toolkit, which is a powerful Python Library that is designed to deal with human language data, especially in natural language processing to perform tasks such as tokenization, parsing, stemming, etc. Eventually, through using NLTK package in Python, there are 6 steps that this research used to preprocess each comment from the post on Reddit:

Step 1 is case nomalization. In this case, all words are converted to the best case. Some might do this step by transferring all words to lowercase, but this research will use 'true casing', which takes the type of case into account. For instance, the sentence "I Really Love Toronto" will be transferred to "I really love Toronto".

Step 2 is filtering out non-essential characters. This step is to remove elements that do not help to have a better understanding in the sentiment in the text, such as punctuation, emoji, as well as non-English characters.

Step 3 is tokenization, which also known as sentence to words breakdown. It splits sentences into words, which are also called 'Tokens'. For instance, the sentence 'I love Toronto too much' will be separated to strings: 'I', 'love', Toronto', 'too', 'much'.

Step 4 is is stop words removal, which also means eliminating common words like 'a', 'an', 'as', 'and' 'are', 'but', 'be', 'for', 'such', 'their', etc. The reason to remove them since most of them are

DOI: 10.54254/2754-1169/102/2024ED0081

meaningless in a sentence, which are not provide any valuable concept or knowledge [6]. Therefore, this step is beneficial to improve the efficiency in the further preprocessing and analyzing steps.

Step 5 is lemmatization, which is also called morphological analysis of words. It involves arranging the various inflected forms of a word so they are identified as one entity, known as the word's lemma or its dictionary form [7]. Overall, this technique is akin to stemming, which involves simplifying a word to its root form by removing its suffixes and prefixes. [8]. Nevertheless, Lemmatization will analyze the words based on its meaning in the sentence. For example, in the case of the word 'better,' a stemming algorithm, such as the Porter Stemmer, might shorten "better" to "bett" or keep it as "better," depending on the specific rules of the algorithm. In contrast, Lemmatization processes the word "better" by understanding it as the comparative form of "good," so it concludes that the base form, or lemma, of "better" is "good". Thus, compared to Streaming, Lemmatization is a better solution to Sentiment Analysis since it is beneficial to get a more accurate outcome at the end of the research.

The last step is reassemble, which means reassembling all separated words back into complete sentences, preparing them for further sentiment analysis.

2.4. Sentiment Analysis

Sentiment Analysis is an area of natural language processing (NLP) that concentrates on assessing the sentiment expressed in text posted by individuals. Through analyzing each word in a sentence, it will identify whether the sentiment of the poster is positive, negative, or neutral.

The primary methods would usually used in sentiment analysis are Lexicon-Based Method and Machine Learning Methods. Lexicon-Based Method relies on a pre-defined list of words(lexicon) that each word in this lexicon has its own sentiment score. In the process of analyzing each segment of the text, this method will be needed to separate the words in the sentences and assigns them specific pre-set scores in the lexicon. Eventually, the sentiment of yhe whole text is counted by the sentiment scores of each word present in the text. On the other hand, Machine Learning approach required sufficient data to train models on labeled datasets to learn how the sentiment is typically expressed [9]. The training allows models to learn which textual features (such as words, phrases, syntax, and structure) correlate with specific sentiment outcomes. Eventually, After comparing the pros and cons of these two methods, Lexicon-Based Method has been chosen as the primary method that use in this study since it does not require the user to have a large amount of data for training and is also straightforward to implement and understand.

Additionally, Textblob is the main tool that used to analyze the sentiment of the sentence in the post. TextBlob is a Python package that offers a simple API for performing typical natural language processing (NLP) functions, such as part-of-speech tagging, extracting noun phrases, analyzing sentiment, classification, and translation [10]. For the sentiment analysis function in TextBlob, it has its own built-in lexicon, and this lexicon includes many words along with their corresponding sentiment polarity values, in range between 1 to -1. For the meaning beehind the number, Negative values in this context suggest a more negative or pessimistic sentiment, with smaller numbers indicating stronger negativity; Zero values represent neutrality, implying neither positive nor negative sentiment. Subsequently, positive values signify a more positive or optimistic sentiment, where larger figures denote greater positivity.

3. Result

Ultimately, after the data collection, there are 490 comments that has been collected. Noticeably, all comments in this post are posted between March 21 and March 29 in 2024, and 99% the comments were posted in the date range of March 21 to March 23, which imply that the result might only be

DOI: 10.54254/2754-1169/102/2024ED0081

available to predict the results accurately for short term. Then, after the collection, the sentiment analysis has been conducted and each comment have been given their specific value. as what mentions in Method, the value is in the range in 1 to -1. If the value is positive and more near to 1, it implies that the user has an optimistic sentiment. Reversely, if it is negative and closer to -1, it means the user is pessimistic to the information in the post.

Subsequently, by counting the sentiments of all comments, the Table 1 has been generated and shown the result below:

Table 1: Statistics of the result of sentimen analysis

Sentiment category	Count	Percentage
Neutral	212	43.27%
Postive	196	40.00%
Negative	82	16.73%

In the statistics and sorting the comments to 3 different categories, the content below shows the range:

- Positive: 0.05 - 1.00
- Neutral: -0.05 – 0.05
- Negative: -1.00 - -0.05

The analysis of user comments reveals distinct sentiments toward the post: 212 comments are neutral, 196 express positivity, and 82 are negative. Notably, a significant 83% of commenters display either a neutral or positive attitude. This substantial majority suggests an openness and optimism toward the subject of the post, which in this context, relates to investing in Reddit stock. It is well-documented that emotions significantly influence human decision-making, especially in financial investments [11], and the investors who have a rich experience might have a more intense emotion that will be express than the inexperience investor [12]. Therefore, in this case, there are a large portion of the reactions are optimistic, so it can be inferred that these individuals may be more inclined to consider investing in Reddit stock, driven by their positive emotional responses.

In order to prove the statement, the Figure 2 from StockCharts.com is shown below and will be used as a tool to do the comparison and evidence:

Figure 2: Actual Movement of stock Reddit

DOI: 10.54254/2754-1169/102/2024ED0081

As mentioned before, most comments are posted between March 21 to March 23. In the chart, the stock trend following this period also strongly supports our point of view as being correct. On March 20, Reddit had there IPO at that day and the stocks fluctuated in the range between 46 and 50 in the next 2 days. However, this circumstance has alternated after the first weekend after the IPO. Since the market opened on March 25th, Reddit's stock underwent a rapid raise, and reaching its peak on the March 26 at $65. Therefore, this is a great evidence to prove that the individual's sentiment might affect their decision in investment. Nevertheless, after the stock reached the peak, it began to fall until it stabilized at $45.97 on April 1st. Therefore, this situation also proves that only 1- or 2-days data might be only available for the prediction to the movement of the stock in short term. However, for the long-term prediction, collecting more comments published over different periods is necessary.

4. Conclusions

Human beings are creatures easily affected by their sentiments, especially when making investment decisions. Meanwhile, with the development of technology, more and more people would like to express and share their opinions and thoughts online. Collecting and analyzing these comments can serve as an effective tool and predictor for understanding public sentiment and market trends in depth. This research indicates that when investors are optimistic or believe that the stock market is bullish, they are likely to enter the market, which might contribute to the growth of the stock price. Conversely, if investors are pessimistic, they might prefer to remain inactive, sell out, or short sell the stocks. Thus, under these circumstances, web scraping and sentiment analysis are valuable tools for collecting people's comments online, analyzing emotions related to specific events, and gaining a deeper understanding of their investment habits and decisions. Nevertheless, in this research, due to a lack of data, only one or two days' worth of data could be collected. Although this helps investors determine the trend for the next few days, it is not sufficient to maintain the same accuracy in the long term. Therefore, future research should aim to collect a large amount of data from different time zones to enhance the accuracy of the analysis. Additionally, establishing and utilizing an emotional analysis dictionary specifically designed for investment analysis would also be a worthwhile endeavor in further research.

References

[1] *George, J. M., & Dane, E. (2016). Affect, emotion, and decision making. Organizational Behavior and Human Decision Processes, 136, 47-55. https://doi.org/10.1016/j.obhdp.2016.06.004*

[2] *Nygren, T. E., Isen, A. M., Taylor, P. J., & Dulin, J. (1996). The influence of positive affect on the decision rule in risk situations: Focus on outcome (and especially avoidance of loss) rather than probability. Organizational Behavior and Human Decision Processes, 66(1), 59-72. https://doi.org/10.1006/obhd.1996.0038*

[3] *Wikipedia contributors. (2024). Reddit. In Wikipedia, the free encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Reddit*

[4] *Reddit Inc. (2024). Press. Retrieved from https://www.redditinc.com/press*

[5] *Dean, B. (2024). Reddit user and growth stats (2024). Backlinko. Retrieved from https://backlinko.com/reddit-users*

[6] *Sahu, S., Divya, K., Rastogi, N., Yadav, P. K., & Perwej, Y. (2022). "Sentimental Analysis on Web Scraping Using Machine Learning Method". Journal of Information and Computational Science (JOICS), ISSN: 1548-7741, Volume 12, Issue 8, Pages 24-29, August 2022. DOI: 10.12733/JICS.2022/V12I08.535569.67004.*

[7] *Khyani, D., Siddhartha, B. S., & Niveditha, N. M. (2021). "An interpretation of lemmatization and stemming in natural language processing". Journal of University of Shanghai for Science and Technology. Retrieved from https://www.researchgate.net/publication/348306833_An_Interpretation_of_Lemmatization_and_Stemming_in_Natural_Language_Processing.*

[8] *Pradana, A., & Hayaty, M. (2019). "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts". Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, ISSN: [missing], Volume 4, Issue 4, [Page numbers missing], 2019. DOI: 10.22219/kinetik.v4i4.912.*

DOI: 10.54254/2754-1169/102/2024ED0081

[9] *Pandya, S., & Mehta, P. (2020). "A review on sentiment analysis methodologies, practices, and applications". International Journal of Scientific & Technology Research, Volume 9, Issue 2.*

[10] *Abiola, O., Abayomi-Alli, A., Tale, O. A., Misra, S., & Abayomi-Alli, O. (2023). "Sentiment analysis of covid-19 tweets from selected hashtags in Nigeria using vader and text blob analyser". Journal of Electrical Systems and Information Technology, Volume 10, Issue 1. https://doi.org/10.1186/s43067-023-00070-9*

[11] *Lad, C., & Tailor, H. (2016). "An empirical study on emotional bias affecting investment decisions of investors". Global Journal of Research in Management, Volume 6, Issue 1. Retrieved from https://www.utu.ac.in/DManagement/download/June%202016/4.pdf*

[12] *Seo, M. G., & Barrett, L. F. (2007). "Being emotional during decision making—good or bad? An empirical investigation". Academy of Management Journal, Volume 50, Issue 4, Pages 923–940.*