Analyzing Social Network-Reddit for

Investment Decisions

## Web Analytics

## Instructor: Professor Jenny Gong

## Fordham University

## 18 December 2022

## Team members:

Kaung Nyunt Paing

Wanpeng Xu

Danying Lin

Jeremy Montoya

ZiyinYue

## Problem Background

Social media and/or social networks have come to be the primary source of information and idea sharing nowadays. So much so that many companies find themselves paying influencers, creating social media marketing departments to connect with customers and gathering data from social media to use in research and development. Well-informed individuals keen to manipulate the market, posted information on social media forums such as the Reddit chatroom r/wallstreetbets, which boasts 4.8 million members, encouraging and rallying the masses to buy shares with the hope that it would drive the price of certain stocks up. Motives behind this onslaught ranged from making personal profit to a desire to squeeze the short positions of hedge funds.

Prompted by the information posted on social media, retail investors began buying these “meme-stocks” including GameStop, AMC Entertainment, Blackberry, and Nokia. The activity sent their prices soaring, with the GameStop share prices climbing over 1000% in just two weeks. The aftermath of using social media to drive the stock price higher and created highly volatile stock and had the intended consequence of financial institutions losing large sums of money. These financial institutions were betting against the stocks, or shorting, betting the stock would lose value while reddit users continued to bet for the stock would gain value, or place calls.

* At the center of this stock trading frenzy was a battle between amateur investors and multi-billion dollar hedge funds. Inspired by this event, we tried to find actionable insights from social networks (reddit for our project) by analyzing the contents discussed on such networks and utilize those insights to help us reach investment decisions on trading stocks and portofolio management.

## Project Overview

It is now generally accepted that the information circulating in the large online forums, such as Reddit and Twitter can influence the financial markets.

*Project Goal:* The goal of our project is to investigate whether the information circulating or contents being discussed on the social media/networks can be used effectively as a reliable source of information for investing in financial markets.

*Procedure:* The step by step process flow of the project is as follows.

1. The collection of data from Reddit
2. Sentiment Analysis on the collected data
3. Evaluate the sentiment scores for stock selection and trading
4. Stock trading strategy and portofolio managment
5. Summarize profit and loss due to selected stock trading for evaluation

*Source and Budget:* The source and budget of the project is as follows.

* The social network being analyzed is solely Reddit. The subreddit being scraped for content mining is ‘Wallstreetbets’.
* The imaginary initial capital for trading stock based on the sentiment analysis is $10,000.

*Timeline:* The project ran for one week. The starting date of data collection is 5th December 2022 and the ending date is 12th December 2022.

## Dataset Collection

For the data collection process, the tool primarily in use to program the content scraping process is Python. It is used in coordination with two APIs - PRAW (Python Reddit API Wrapper) and/or Pushshift API. The program also uses the Pandas library.

To understand the step by step data collection process which comes later, it is important to give some description on how Reddit and particularly, the sub-reddit “wallstreeetbets” operates.

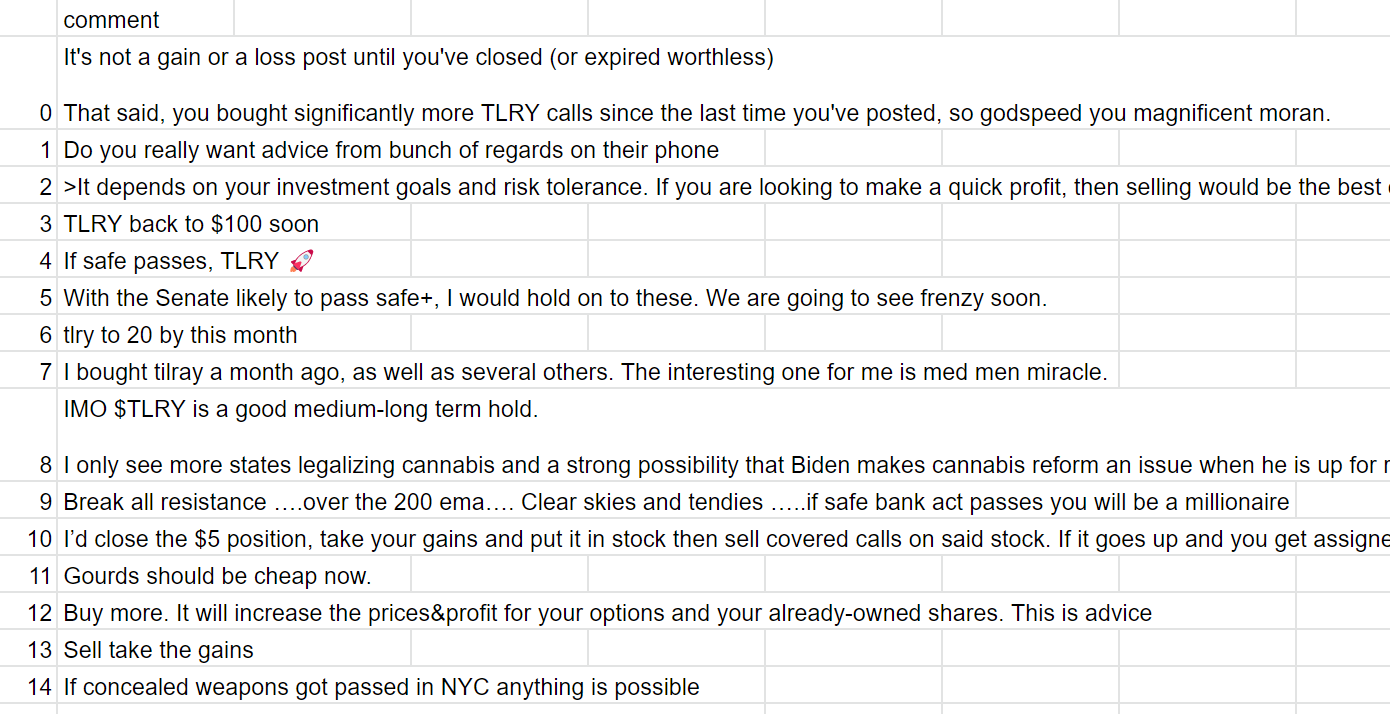
* The topic or thread or post that someone posted for discussion will have a score (similar to how tweets in twitter works). This score will be based on the upvote (equal to +1 score) and the downvote (equal to -1 score) that other users give to the post. The higher the score, the more popular the thread of discussion will be.
* Each post will have a title. Users can also comment on the posts.
* In reddit, you can see the top posts (the most popular posts according to score) in a specific sub-reddit in terms of today, this month, this year and all time.
* For sub-reddit “wallstreetbets” , when users are discussing about a specific stock, the thread or post title usually have the sign ‘$’ in it.

The detailed step by step data collection process is as follows.

1. For the first step, we want to collect the top posts being discussed on the sub-reddit “wallstreetbets” on that day. We wrote a Python program utilizing the PRAW to get the top posts in the sub-reddit along with each post's title, the url of the post , score of the post, and the number of comments the post has.
2. Secondly, we need to pick posts that are discussing stocks out of all the posts collected. We wrote the program to filter the posts by going through the titles of each of the collected posts and find the word that start with ‘$’ using a lambda function and selecting only the posts with ‘$’.
3. Among the filtered top posts, we picked the top 3 posts (highest scores) of each day of the project for further investigation of analyzing the sentiment of the comments of each post.
4. Finally, using the url of the each of the selected posts, the program will get the comments of the posts using the PRAW again and put the comments of each posts into CSV files using Pandas library.

Note: The process is based on the Python program using PRAW. If Pushshift API is used to collect the data, since the API does not have top posts filtering function. If Pushshift API is used, you will have to manually select the time frame for data scraping and then sort the posts in terms of score to get the top posts.

The data collection Python program ran at 4 pm everyday of the project week. The idea behind setting 4 pm to run the program is because the official stock market closing time is at 4 pm. We want to trade stock right before the market closing time using the sentiment analysis based on the comments of the thread discussing the stocks on that day. Otherwise, the sentiment analysis of the stock discussion are irrelevant; investment decisions using the sentiment score of the stocks being discussed that day should be on that same day and not the next day.



(Figure 1 Comments of Posts discussing stock example)

The following table shows the selected stocks for each day of the project timeline.

| **Day** | **Stock No** | **Name** | **Full Name** |
| --- | --- | --- | --- |
| 1 | 1 | AAPL | Apple |
| 2 | CVNA | Carvana |
| 3 | TLRY | Tilray Brands |
| 2 | 1 | TSLA | Tesla |
| 2 | TSMC | Taiwan Semiconductor Manufacturer |
| 3 | BTU | Peabody Energy Corporation |
| 3 | 1 | CVNA | Carvana |
| 2 | CHGG | Chegg |
| 3 | AAL | American Airlines |
| 4 | 1 | TSLA | Tesla |
| 2 | SPXS | Direxion Daily S&P 500 Bear 3X Shares ETF |
| 3 | ORCL | Oracle Corporation |
| 5 | 1 | SPY | SPDR S&P 500 ETF Trust |
| 2 | AAPL | AppleOracle Corporation |
| 3 | LI | Li Auto Inc |

(Table 1 15 Most Mentioned Stocks )

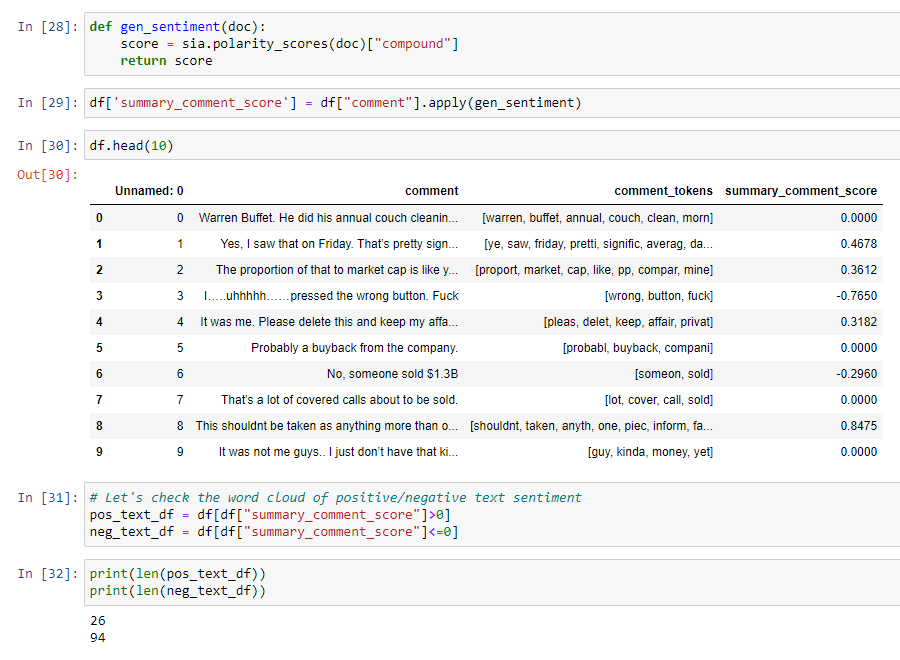
## Sentiment Analysis Process

In order to analyze the comments obtained from data collection process, we used the python NLTK library for extracting the sentiment of the comments mined.

* The first part of our analysis by creating tokens and removing stop words from the list of comments we previously had extracted from Reddit. This was a several layer process and will be included as a separate section.
* After the cleaning and processing of the data, we create a module to process the comments and gain the sentiment scores. This module was applied to the comments and separated into two different lists of positive and negative comments.
* The count of the number of positive comments and negative comments of each post are aggregated to check the majority.



(Figure 2 Text Data Preprocessing)



(Figure 3 Sentiment Analysis)

Chart, bar chart

Description automatically generated

(Figure 4 Aggregate of the count of positive and negative comments)

## Stock Trading Strategy

The strategy to either buy or short a selected stock is very straightforward.

* If the number of positive comments for the post or thread discussing the selected stock is majority, the strategy is to buy that stock.
* If the number of negative comments for the post or thread discussing the selected stock is majority, the strategy is to short that stock.

## Portofolio Management Strategy

In order to hedge the risk of our portfolio, we assigned the portion of our portfolio by mimicking the position of indexed ETFs. We will invest a specific portion of our fund in a specific equity based on which indexed ETFs hold the stock. The specific portions are determined by the risk of the index that the indexed ETFs are tracking. For example, for stocks that are holding by ETFs that are tracking S&P 500, we assign 30 percents of our portfolio to that stock because S&P 500 index is considered low risk index; for stocks that are holding by ETFs that are tracking Nasdaq, we assign 25 percents of our portfolio to that stock because Nasdaq index is consisted of high-tech companies which have relatively high risk. Moreover, for stocks that are held by two or more ETFs that are tracking different indexes, we used the average number of the portion. For example, APPL is held by both ETFs that are tracking S&P and Nasdaq. we assigned 27.5 percent, which is the average of 25 and 30, to AAPL. Furthermore, because investing directly in the indexed ETF is considered low return investment, we only assigned 20 percent of our portfolio to them. Lastly, because those leveraged indexed funds are of high risk as they are already leveraged, we assigned a 20/ leveraged ratio of our portfolio to them. For example, SPXS is 300 percent leveraged against S&P 500, we assigned 10 percent of our portfolio to it. The specific portions of each indexed ETFs are listed below:

| **Portfolio** |  |
| --- | --- |
| **Index** | 20 |
| **Leveraged Index** | 20/leveraged ratio |
| **SPY** | 30 |
| **QQQ** | 25 |
| **DJIA** | 30 |
| **IWM** | 20 |
| **OTC** | 10 |
| **other** | 5 |

## Results & Evaluation

The data yielded some great results and insights into the markets and the effectos of social networks.

As previously mentioned, our goal was to look at the sentiment if the scores were more positive it was in our best interest to bet for the stock, whereas if it was negative our best interest was to short the stock.

Chart, bar chart

Description automatically generated

(Figure 4 Aggregate of the count of positive and negative comments)

It can be generally seen that there is an overwhelming amount of negative comments compare to positive comments leading to the aggregate score of each posts leaning towards the right. (Negative)

* The first day of pulling comments, our stocks to look at for that day were Apple, Carvana and Tilray. The comments we had pulled returned a negative sentiment score, which in turn helped us decide to short the stocks for the day.
* The second day, our stocks to look at were Tesla, Taiwan Semiconductor Manufacturing Company (TSMC), and Peabody Energy Corporation(BTU). The sentiment scores for Tesla and TSMC were negative which gave us cause to short the stocks, although BTU did have positive sentiment score which leaned more positive, we decided to also short this stock that day.
* The third day brought us the stocks of Carvana, American Airlines, and Chegg. American Airlines and Carvanna had negative scores that were high, which ended up being shorted, while Chegg had positive scores and led us to bet for the stock that day.
* The fourth day of data found us looking at the stocks of Tesla, Direxion Daily S&P 500 Bear 3X Shares ETF (SPXS), and Oracle. These sentiment scores for Tesla and SPXS were again more negative, while Oracle was more positive which again caused us to react in the manner we had been doing, shorting for more negative sentiment scores and betting for stocks with more positive scores.
* The final day of our project was pulling information again for Apple, SPDR S&P 500 ETF Trust (SPY), and Li Auto. Apple and SPY we again found ourselves betting against because of negative sentiment scores, while Li Auto we found ourselves better for as the sentiment scores were more positive.

Below we have included a table of our stocks and revenue earned or lost on the day based on actual data from stocks previously mentioned.

| DAY 1 | | DAY 2 | | DAY 3 | | DAY 4 | | DAY 5 | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AAPL | $69.7 | TSLA | $121.5 | CVNA | $(175.28) | TSLA | $(113) | SPY | $ (33) |
| CVNA | $27.4 | TSMC | $3.5 | AAL | $128.9 | SPXS | $(19.2) | AAPL | $ 51.6 |
| TLRY | $66.2 | BTU | $-47.8 | CHGG | $43 | ORCL | $(9.9) | LI | $ (35.9) |

The results yielded $77.79 of revenue from the week or approximately .779% growth.

(Figure 5 Waterfall chart of profit and loss)Chart, waterfall chart

Description automatically generated

## Conclusion and Recommendation

Overall, the results from our project suggest that using information from social networks would be a great starting point for retail investors and hedge funds to formulate ideas for new investment projects.

However, we realized due to the short project timeline for data collection leading to lack of abundant data, we were not able to reach a comprehensive conclusion of whether this practice can be used reliably or not.

Furthermore, we concluded that basing solely on information from social networks and/or social media for investment decisions such as stock trading is not viable. Thus, using different sources including information from social media as a factor for decisions is recommended.

For future projects, we suggest a larger time frame and adjusting of the NLTK python tool to accomodate the bias of the content from the social network being researched (e.g. Data suggest that reddit content is more negative than positive).