Reddit & Stock Market Data Analysis

Aneesha Sreerama

Northeastern University, Boston, MA, USA

Introduction

The objective of this analysis is to find correlations between the movements of the stock market with the content on Reddit's largest investment community, r/wallstreetbets. Data was collected using the Reddit and the Yahoo Finance APIs. The goals of this project included successfully scraping data through Reddit, using natural language processing to analyze the most frequent words, observing correlations between the number of times a company was mentioned and its price movement, determining the ratio of the number of times "call" and "puts" were used over the course of 2020, and using regression models to better analyze the dataset. The motivation and significance of this analysis was to better understand the social media response to financial gains and losses. Moreover, a goal was to determine if this data source could be used for predictive modelling as this data source is often used by quants [2]. The driving hypotheses for this analysis included that the Calls-to-Puts ratio would increase over the course of the year, that the most talked about companies this year would include Tesla and Apple, and that there is a negative correlation between the price of the SPY ETF and the number of times "Puts" was mentioned in the subreddit.

Data Sources and Methods

I acquired my data using the Reddit's PRAW (Python Reddit API Wrapper) [1]. I created an account and modified the settings in order to attain a client ID and a client secret code. I used these codes to access PRAW and create a reddit object. Using the documentation, I accessed the r/wallstreetbets subreddit and changed the settings to see all the top posts in 2020. For every submission I extracted the time created in UTC format, the comments on the post, the number of upvotes the post had, the tile of the post, and the ratio of upvotes to downvotes. To obtain the financial data I used the requests and pandas data reader module in order to read financial data for a given ETF or company for the specified timeframe.

In order to clean the data into a useable format I first had to convert UTC time into a Datetime format. In order to do so, I converted UTC time into a string format in the Pacific Time zone. I parsed the data using the datetime module in order to convert the value into a Datetime format. Finally, I utilized the time delta module to change my time into Eastern Daylight Time.

In order to process the comment data, I had to clean the characters used in the comments. The comments were littered with punctuation, extra spaces, emojis, and links. I parsed through the dataset to remove all the unwanted characters and stored the cleaned in a new column in my dataframe. To further clean the dataset, I removed the English stop words and split the long string into a list of words.

The financial obtained from Yahoo Finance was clean and organized when I accessed the data through their API. However, in order to merge dataframes I had to change the index of the dataset (the dates). The dates were in Timestamp format whereas the rest of the dates were kept in string format. In order to make the dataset compatible for merging, I reset the index of the dataframe and converted the dates from the Timestamp format to string format using type casting, string splicing, and list comprehension.

Use Cases

The ways for the user to interact with this code include

1. A function to scape reddit data (a user can specify the subreddit name and the number of posts they would like)
2. A function to cleans all the comment data by stripping punctuation, extra spaces, emojis, and links given a string
3. A function to create a word clouds for any given month
4. A function for retrieving stock data given a company/ETF, start date, and end date
5. A function for plotting the number of mentions of company on a subreddit versus its share price in the given timeframe
6. A function for plotting the number of mentions of any give word vs the SPY ETF or ^VIX ETF.
7. A function that renders a graph listing the Top 10 words used in the posts for a specific day (given the day)
8. A function that plots the daily Open, High, Low, and Close for the financial data of any given company/ETF

Analysis and Results

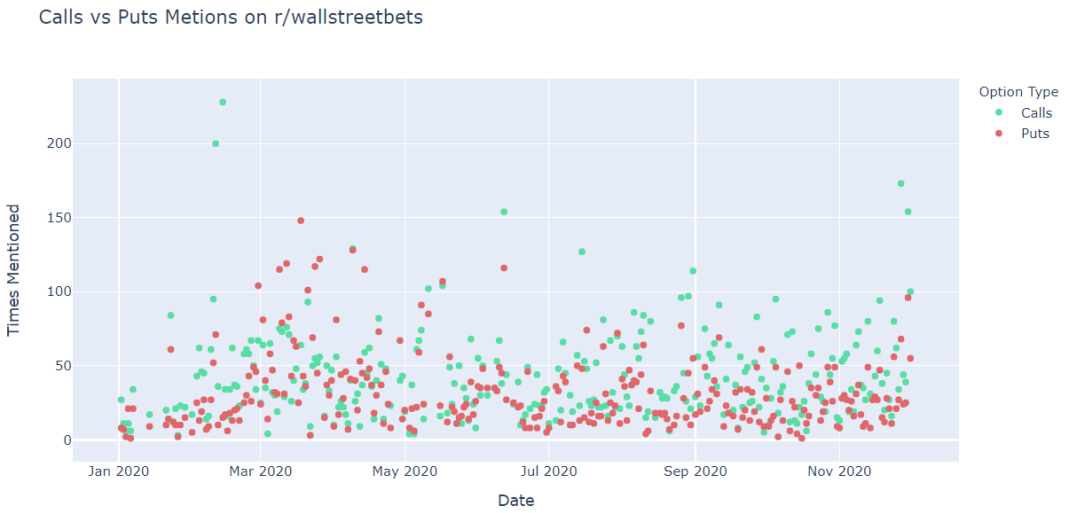
After extracting, cleaning, and processing the original data, I wanted to have a better understanding of the dataset and an overview of the few hundred posts that were collected from the subreddit. In order to do so, I decided to make a word cloud for every month for 2020. This would provide a rough overview of what the highlights were for that month.

The two things that were most notable for me included the following

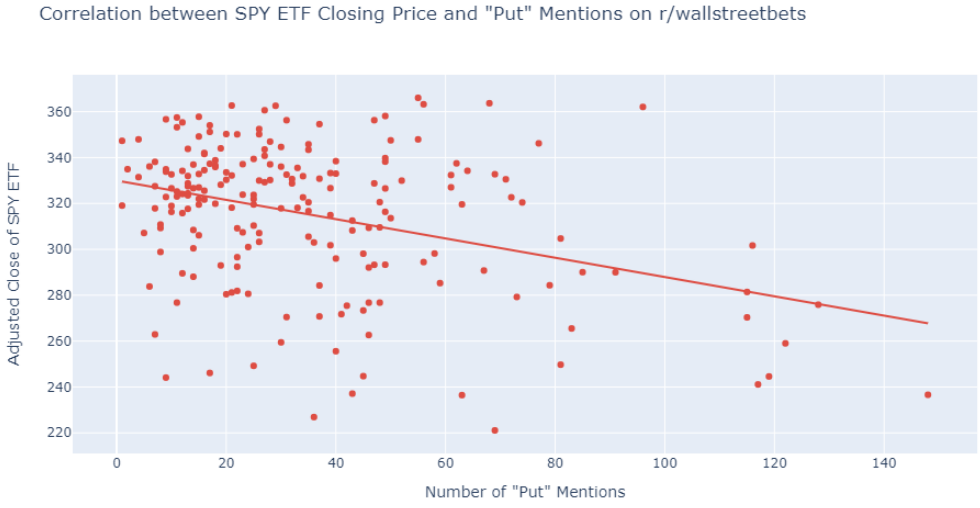
1. The sizes of the word “call” and “put” seemed to vary from month to month. Calls and Puts are two different options you can buy in the stock market. A call option implies that the buyer believes that a certain company is going to see major gains in a relatively short period of time. A demand for call options implies that the public believes that massive gains are to be made. On the other hand, a put option implies the very opposite: the buyer believes a company will see major losses. As you can see above, “call” was used a lot in January. However, as the year progressed the words get used less often up until August. The word becomes more popular over the second half of the year, particularly between September-November.
2. Some of the most talked about companies for the year included Apple and Microsoft in February. In addition, Tesla, Nikola, and Palantir were topics of discussion towards the

end of the year. I wanted to investigate what may have caused these companies to become a “hot” topic of discussion in these time frames.

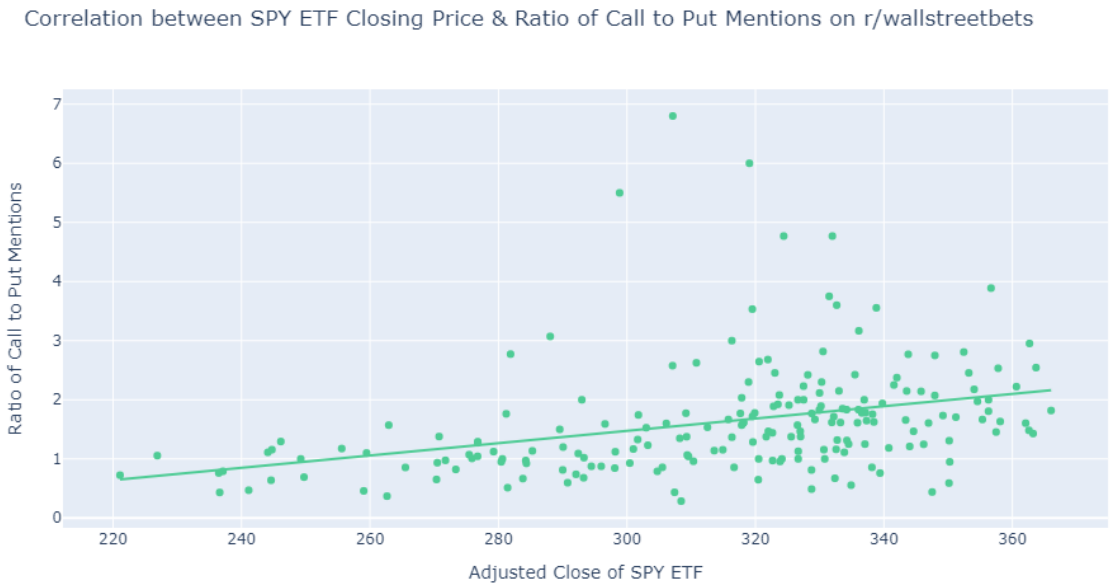
I first plotted the number of times “call” and “put” were mentioned on the subreddit over the course of the year. Each day has two data points corresponding to it: number of calls in blue and number of puts in red. As you can see the number of people who were discussing “puts” tended to be greater in the first half of the year and the number of “calls” tended to be higher in the second half of the year.



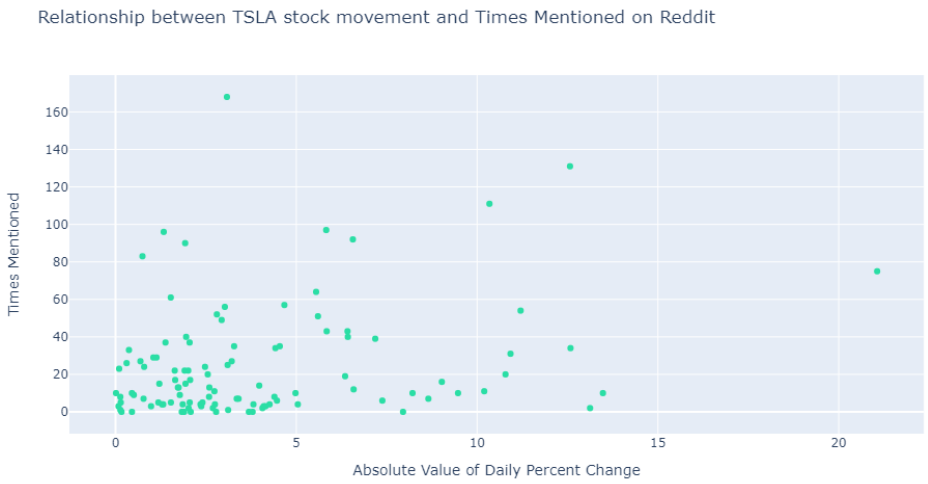
To take a deeper look into this pattern, I plotted the number of times “put” was mentioned with the adjusted closing price of the SPY ETF. The SPY ETF is an indicator of the health of the market as it considers the share prices of the Top 500 companies. There is weak but negative correlation between the “Put mentions” and the closing price. The general trend suggests that a negative social media response is associated with an unhealthy market.



In order to affirm this general trend, I wondered if there would be a positive relationship between the Adjusted Close of the SPY ETF and the ratio of “calls” to “puts”. The idea was that if ratio of “calls” to “puts” was higher, people must believe that they think that the market would see future gains. In addition, this is associated with a healthier market. The scatterplot shows a weak but positive association with the Adj. Close and the Ratio of Call to Put Mentions. Moreover, the sheer number of outliers in the dataset are contributing to a low R^2 values and using a trimmed dataset could have produced a better result. However, this trend and the one seen before do suggest that there is a relationship between the online sentiment and the performance of the stock market.



The word clouds also hinted at looking into how many times a company was mentioned and why it may have caused surges of popularity. I investigated Tesla’s data as it experienced significant positive and negative changes throughout the year. After plotting the absolute value of percent change and the number of times the company was mentioned, there seemed to be no clear correlation between the two variables. Some factors that may have resulted in this result include that in times of large changes, most of the other companies were experiencing a similar change as COVID-19 impacted the whole community. In addition, as a company’s share price becomes more volatile, significant change seems to draw less attention over time.



The last aspect of the subreddit that I wanted to analyze was if there were certain words that related to the general trend of either the S&P 500 or the Volatility Index. When testing out groups of keywords, the words seemed to follow their own pattern vastly unrelated to the either index. This is an example of one category of words that were used and there seems to be no visible correlation between the variables. This analysis was too general to produce any results.

However, in order to better understand the most poplar words to analyze I made a function that plots the most popular words used by the day. The following is an example of what the function generates.

After running this function over a series of days, it affirmed the original result that the popularity of certain words seems to follow a pattern unlink the general market or the volatility index.

Conclusions

One of the goals of this was to use natural language processing in order to analyze the most frequent words. The word clouds were an excellent visualization technique that summarized the vast amount of data on r/wallstreetbets. I originally hypothesized that the Calls-to-Puts would increase over time. The results of my findings did show that this was in fact true and I went on to find that there was in fact a positive, however weak correlation between the Calls-to-Puts ratio and the closing price of the SPY ETF. Moreover, this trend was affirmed by the negative correlation between the price of the SPY ETF and the number of occurrences of the word “Puts” in the subreddit. The motivation of this analysis was to better understand the social media response to financial gains and losses. I was surprised to see several general trends on social media that do correlate with the stock market.

However, I did see the results of some of analysis failed to provide any significant results. Some patterns I thought I would find between the change in the share price and the company’s popularity failed to produce any significant patterns. The art of finding patterns and the predictability of stock price movement proved to be far more complex than I originally thought, and this was in fact reflected in other part of my data analysis. When analyzing the frequency of key words in trading that were seen in the word clouds did not show any patterns that were in relation to the relative changes of the market.

This analysis is subjected several limitations. Using one subreddit to encompass the opinions of the population of investors produces skewed results as this is not a sample that is representative of the entire public. In addition, as I was looking through only the top posts for this analysis, there could be other posts that reflect different sentiment that I never analyzed. Moreover, there are so many variables that influence the direction of the stock market and I only analyzed one aspect of how the stock market moves. There are a vast number of confounding variables that I have not considered in the scope of the project. In addition, using more powerful

natural processing techniques such as n-grams to better understand the context in which these words were being used in could provide better results. One of the reasons I was drawn to this analysis is because 2020 saw some of the largest changes with respect to the stock market. However, these results may only reflect what happened in 2020 as opposed to a general trend for every year.

References

1. Gan, C. J. (2020, October 19). 5 Quant Strategies used by a Wall Street Trader. Retrieved from https://medium.com/datadriveninvestor/5-strategies-in-quant-trading-algorithms-f4f782d152e2

2. Samrega. (2020, October 07). How Robinhood and Covid opened the floodgates for 13 million amateur stock traders. Retrieved from https://www.cnbc.com/2020/10/07/how-robinhood-and-covid-introduced-millions-to-the-stock-market.html

3. The Python Reddit API Wrapper¶. (n.d.). Retrieved from https://praw.readthedocs.io/en/latest/