# Executive Summary

Reddit Sentiment Stock Analyzer:

This application gathers data on Reddit posts’ sentiment surrounding investments and compares it with stock performance using scatter plots.

The Reddit posts are taken from investing subreddits, especially those geared toward individual investors (e.g., r/WallStreetBets, etc.). Once the application has the data from Reddit, it runs rudimentary sentiment analysis to determine if the posts are positive, negative, or neutral using NLTK Polarity Scores. If more than one post is pulled per day, the application takes the average sentiment score from each day. These sentiment analysis scores are plotted against the date on which they were posted to see how Reddit investors view a stock over time.

The stock performance data is taken from AlphaVantage. We use pd.pct\_change to compute daily returns and plot these numbers over time. We plot the daily returns on the same output as the sentiment scores to visually see how Reddit sentiment corresponds to stock performance. Finally, the application returns the correlation between the sentiment and daily stock returns.

The user must input a valid stock ticker of a publicly traded firm and a desired start date in order to generate the desired output. However, once the application receives appropriate input, it does not need anything else to generate the output.

In the future, this application could be expanded to actually calculate correlations between sentiment and performance, thus being more useful for predicting the next meme stock. The application could also be reworked using an API that calls data from Earnings Calls and analyzes that sentiment, which may be more relevant to stock performance than Reddit posts. Also, the sentiment analysis could utilize a custom dictionary that is geared toward the language of investing to make the polarity scores more accurate.

Usability wise, the application could accept input in the form of company name and find the corresponding ticker symbol on its own. Also, the application could be packaged into a GUI to hide the code and make it simpler for users to follow.

Riley Brady ([rbrady10@villanova.edu](mailto:rbrady10@villanova.edu))

Martin Manion ([mmanion6@villanova.edu](mailto:mmanion6@villanova.edu))

# Technical Overview

The Reddit Sentiment Stock Analyzer application was written in Python in JupyterLab.

### Attributions / Modules

The application owes attributions to the following sources:

The getURLdata function was provided by Professor Sue Metzger.

Although we wrote all of the sentiment analysis code ourselves, we relied on the guidance of [realpython.com](https://realpython.com/python-nltk-sentiment-analysis/). This website introduced us to the NLTK Vader functionality. We utilized the vader\_lexicon package and Sentiment Intensity Analyzer to create polarity scores for each Reddit comment. A potential limitation on this method is that Vader is pre-trained by someone else and does not necessarily have an investing focus. Words like ‘hold’ or ‘‘short,’ which have specific meanings when used in an investment context, may not be properly measured by these polarity scores.

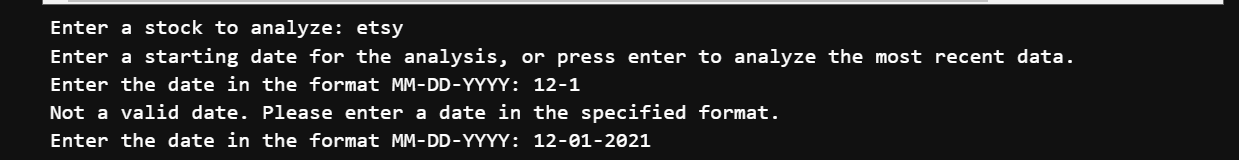
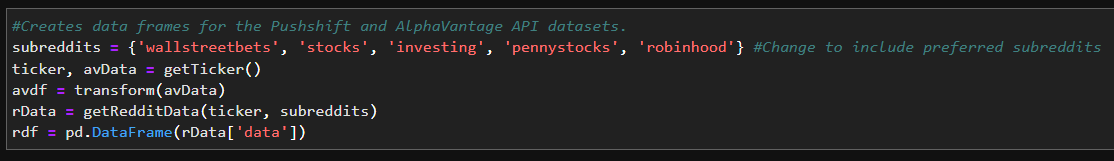
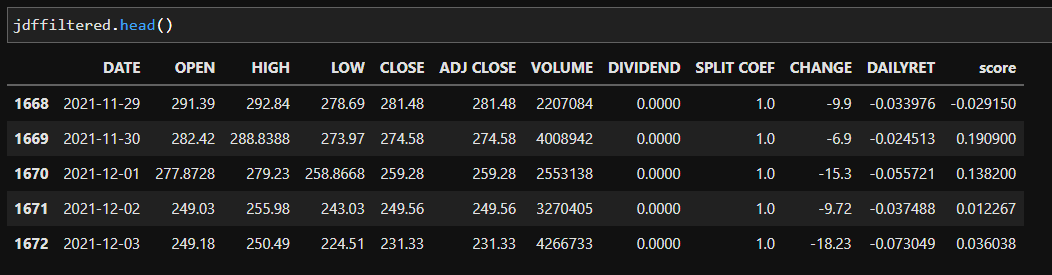
### API Access

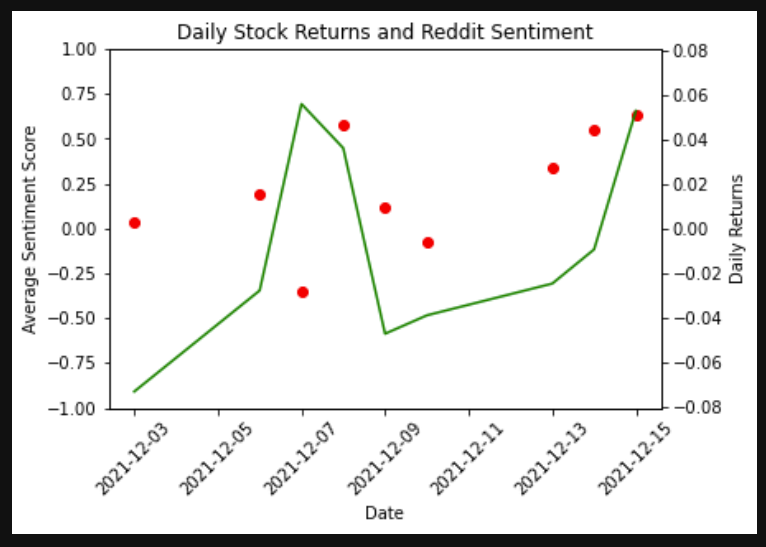
The application is dependent upon use of a APIs providing real-time data. To use these APIs, access keys are issued by the API providers. The following is a list of the APIs on which this application is dependent as well as the required access keys.

Included in the listing is where the access keys can be retrieved, which keys are necessary and the code from this application that will require modification to apply updated keys. In addition, any known constraints about the API have been listed.

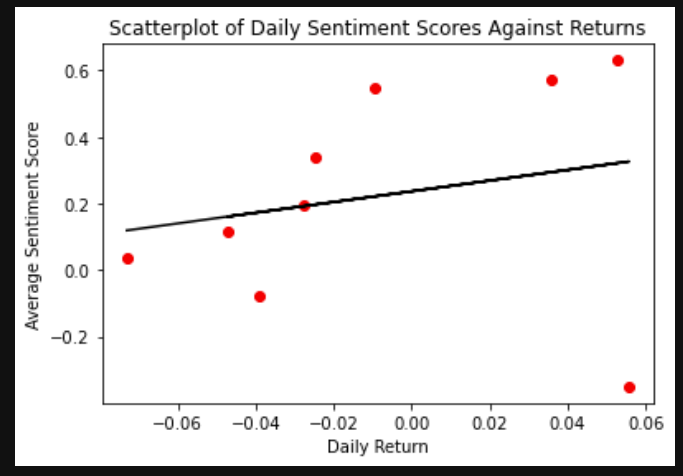
|  |  |  |
| --- | --- | --- |
| Alpha Vantage  Requires an account on <https://www.alphavantage.co/>. [Documentation](https://www.alphavantage.co/documentation/). | | |
| Requires symbol, function, and API key.  If using an updated API key, one must update the getTicker function. | User provides Symbol  We use TIME\_SERIES\_DAILY\_ADJUSTED function.  API Key (set in getTicker function): J9DNBJ9PBMG4O5DW | Limitations on use:   * 5 requests per minute   500 requests per day |
| Pushshift API  No account required. [Documentation](https://github.com/pushshift/api). | | |
| We search specific subreddits for comments and pass a stock ticker as our query and a date for how far back the search goes. | Subreddits are defined in the function. We chose subreddits with an investment focus: r/WallStreetBets, r/investing, r/Robinhood, r/stocks, and r/pennystocks.  We set comments as the search in our function because the comments data is much easier to work with.  Stock ticker and start date are user inputs in our application. | Limitations on use:   * None. |

# Documentation (i.e. end-user training materials with screen shots)

1. First, run the imports and define the functions.
2. Then, run the getTicker function to ask the user for a stock ticker. Input a valid stock ticker.
3. Run the getUserDate function to ask for a start date. Input a valid start date or hit enter for the most recent data.
4. Allow the code to run, it will retrieve the data from both APIs, clean it, analyze the Reddit sentiment, and create a new dataframe with the combined data. 
5. Run the plot statements to see the sentiment and stock performance plotted against time.



1. The second block of plot statements plots the stock performance against sentiment to see any correlation more clearly.



1. Finally, the application also delivers the calculated correlation between the two variables to see the relationship numerically.

