Regression Modeling: A Data Driven Approach to Extracting Insights from Sentiment Analysis and Stock Market Returns

Calvin Li, Nathaniel Yee, Rishpiath Satter

Northeastern University, Boston MA, USA

Abstract

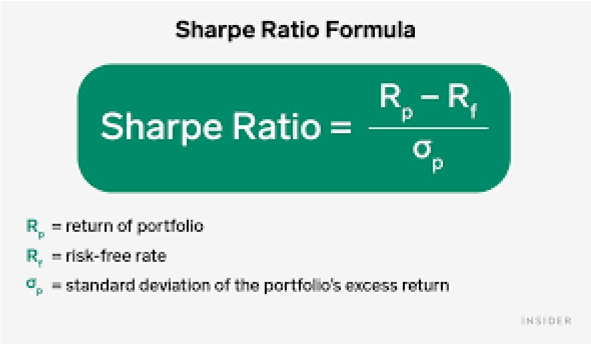
This experiment aims to explore any correlations by running sentiment analysis on data retrieved from Yahoo Finance and the subreddit r/wallstreetbets, and hypothesizing whether or not there is a statistically significant effect between reddit sentiment and stock market risk adjusted returns. The null hypothesis suggests that there is no significant correlation between stock returns and Reddit Sentiment Scores, while the alternative hypothesis suggests that there is a statistically significant correlation between stock returns and Reddit sentiment scores. We also created a similar hypothesis to explore sentiment's connection to time and volatility. Through data preprocessing, sentiment analysis and statistical analysis, the relationships between sentiments, volatility and risk adjusted returns are analyzed.

Introduction

Is there a correlation between financial reports sentiment, reddit sentiment, volatility, and risk adjusted returns? Market sentiment reflects the perception, outlooks and perspectives in which potential patterns can be drawn. In this research, sentiment scores were retrieved from textual data, market volatility was calculated based on historical stock price movements, and levels of risk associated with stock price fluctuations were assessed. Analyzing the relationships between market sentiment and financial information on investment outcomes can be useful to investors, traders and financial analysts in making informed decisions and developing effective investment techniques. The role of financial reporting was taken on for this study of such features, looking at how media sentiments from Reddit and Yahoo impact the volatility and overall risk adjusted returns of the market.

Methods

Our main methodology utilized T-Testing on the correlation coefficients of our variables, a method that takes the two groups we are testing to determine whether the outcomes are statistically significant or likely to have occurred purely by chance. By comparing the p-values returned from the t-testing we utilized an alpha or significance level of .05, as higher significance levels could lead to greater margins of error. Hypothesis testing was utilized to check if there is a statistically significant effect between reddit sentiment, stock market risk adjusted returns, standard market returns, time, and volatility. For context the risk adjusted return was calculated using the Sharpe Ratio as displayed below.



Sharpe Ratio Formula

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

R_p = return of portfolio

R_f = risk-free rate

σ_p = standard deviation of the portfolio's excess return

The null hypothesis was that there is not a significant correlation between stock returns, time, and Reddit Sentiment Scores. The alternative hypothesis was that there is a statistically significant correlation between stock returns, risk adjusted returns, volatility and Reddit sentiment scores. It is important to distinguish that each of these variables were independently tested against the reddit sentiment. When calculating the p-values themselves we utilized the package Scipy and Pearson R to calculate and return our results.

For access to Reddit's API the PRAW package was used to retrieve the top reddit post titles, clean and tokenize the text, and pull out relevant stock information to create large lists on which sentiment analysis could then be performed. [1] To precisely analyze the sentiment of reddit posts we first preprocessed and cleaned the posts using the re package to remove any emojis from the posts using unicodes from the subreddit r/wallstreetbets. As for the sentiment analysis, the Sentiment Intensity Analyzer and VADER lexicon library (Valence Aware Dictionary and Sentiment Reasoner) [2] was run on the reddit posts which was retrieved from the Natural Language Toolkit [3]. Sentiment scores were assigned in a list of dictionaries as either compound, positive, negative, or neutral for assessment.

The data used in this project contains historical stock price data pulled from the Yahoo Finance API which was read into a dataframe for plotting and analysis. More specifically we examined Berkshire Hathaway's 20 holdings and would be considered a balanced portfolio. We set the time frame to be exactly one year from June 19th, 2022 to June 19th, 2023 in order to work through certain limitations that will be later explained. Using the API calls the daily historical stock data for all 20 holdings in the past year into pandas dataframes. [9] All closing price columns were merged based on the date. In this case no data was missing as the historical data is entirely tracked without missing information. Once this was completed, we calculated

[10] daily portfolio returns, the metric volatility, and sharpe ratio, and used it to visualize stock price movements, as well as graph representations of the sentiment trends.

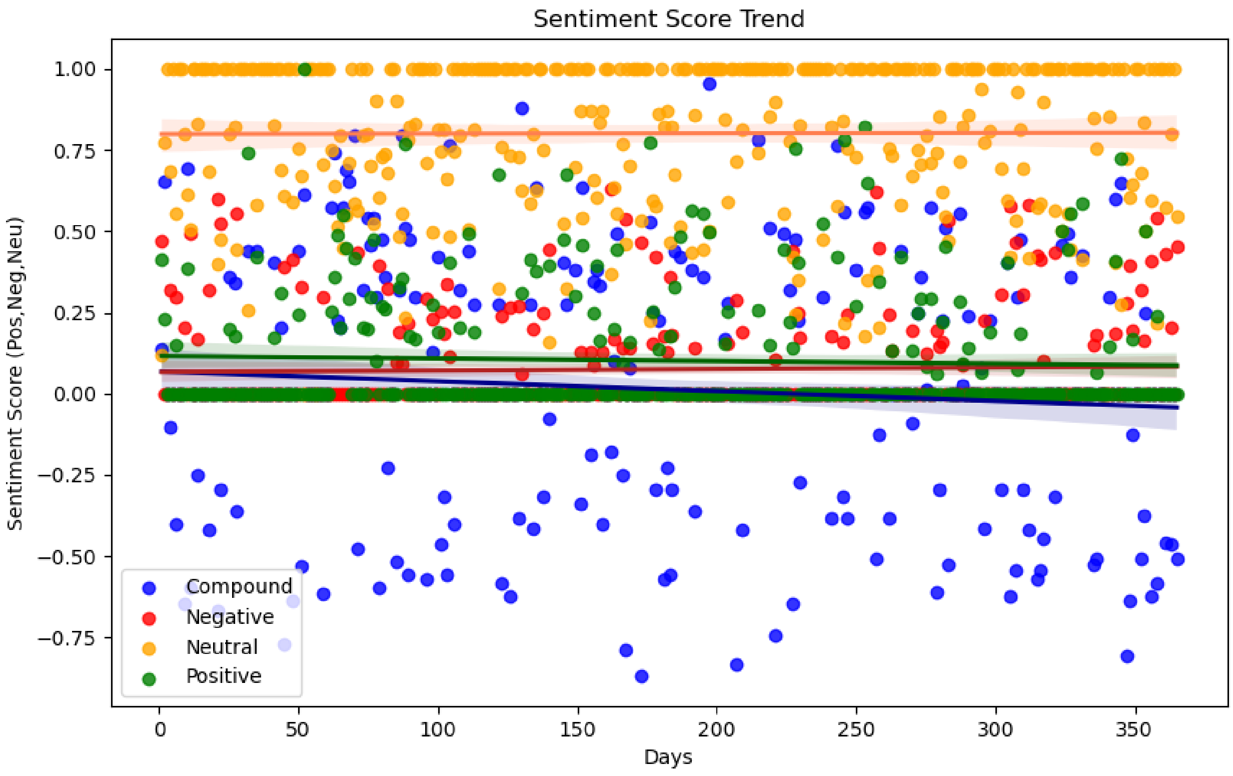
For the statistics and plotting portion of analysis we applied seaborn and scipy. Scipy was utilized to calculate the p-values of the t-testing and would return the resultant of the test based on the alpha of .05 in the form of a printed statement of whether there was a statistical significance. Seaborn plotting was utilized due to the usage of pandas dataframes where regression plots would show the correlation between the variables that we were comparing.

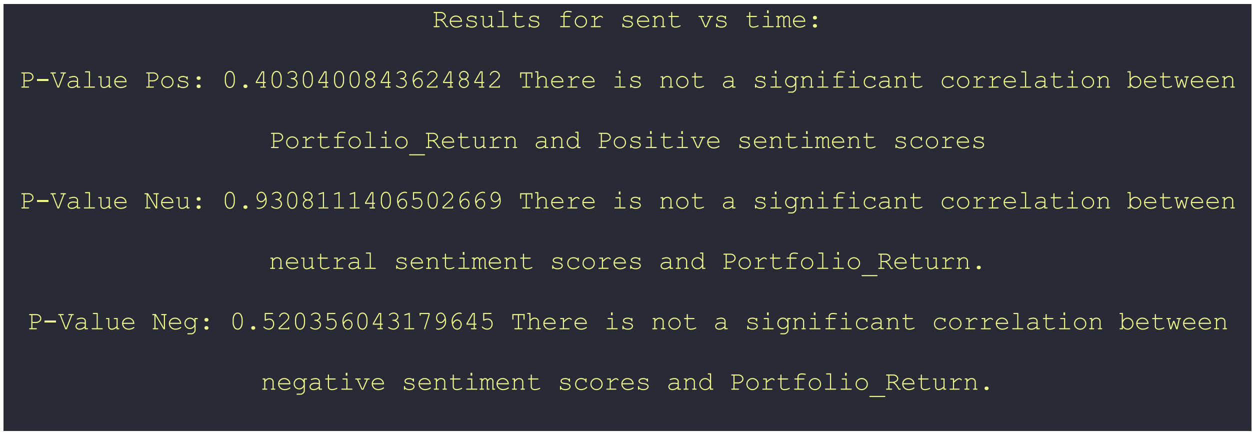
Sentiment scores were plotted over time and plotted over the investment variables as well and visualized as scatter plots to depict the trends and patterns. The statistical and visual representations were analyzed together to draw conclusions on the impact of sentiment on investment returns. For this program the Python libraries that were utilized include pandas, seaborn, matplotlib, nltk, and plotly.

Analysis

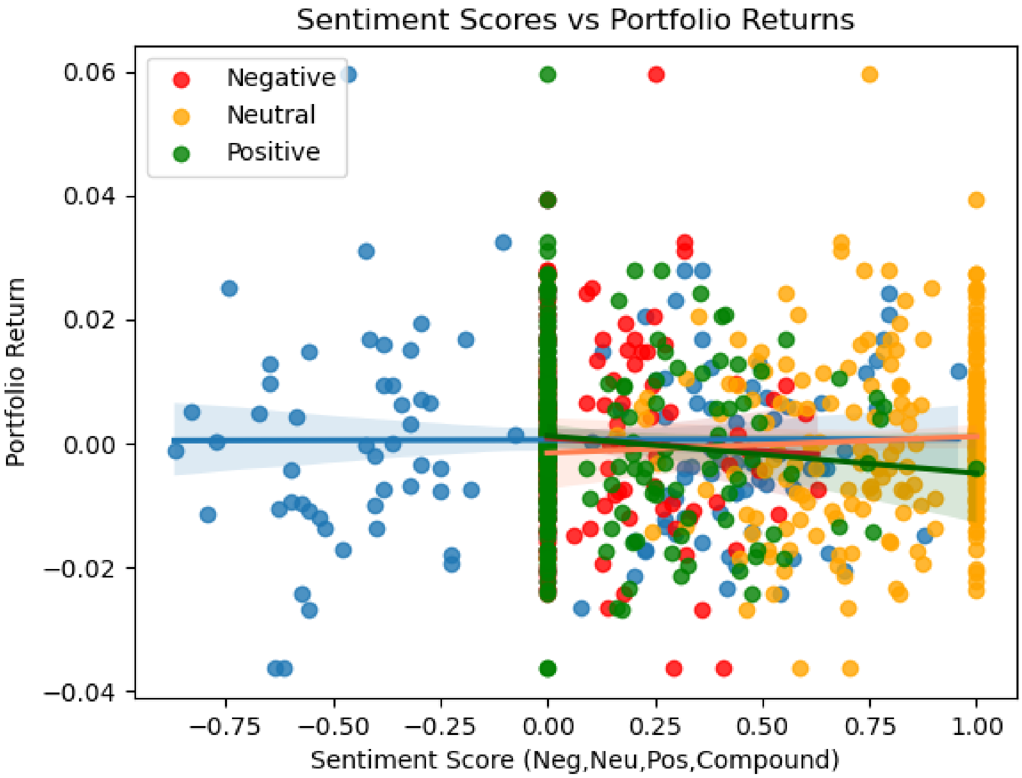
The variables that were measured in relation to the sentiment were sentiment vs. time, sentiment vs. portfolio returns, sentiment vs. sharpe ratio, and sentiment vs. market volatility.

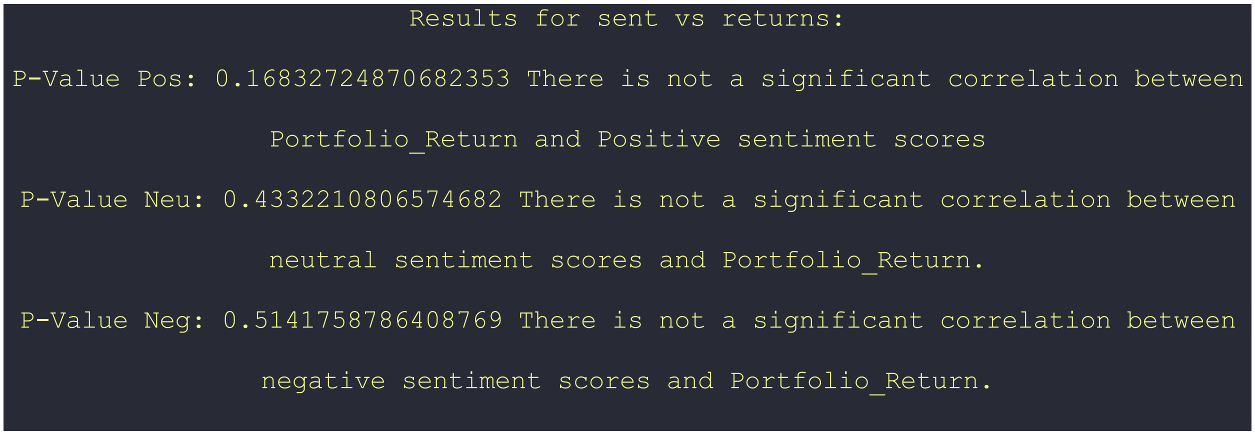
Sentiment vs Time




```
                     Results for sent vs time:

P-Value Pos: 0.4030400843624842 There is not a significant correlation between

                 Portfolio_Return and Positive sentiment scores

P-Value Neu: 0.9308111406502669 There is not a significant correlation between

                 neutral sentiment scores and Portfolio_Return.

 P-Value Neg: 0.520356043179645 There is not a significant correlation between

                negative sentiment scores and Portfolio_Return.
```

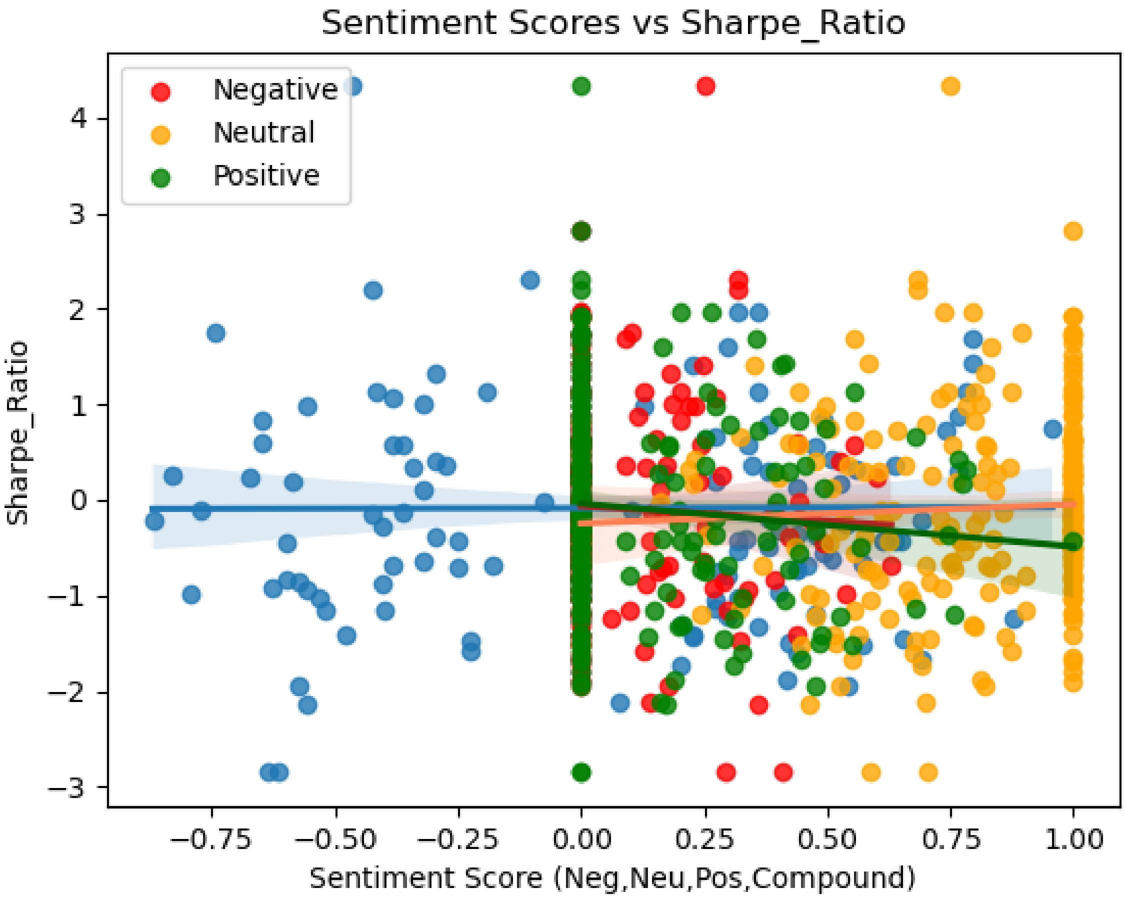
Sentiment vs Returns

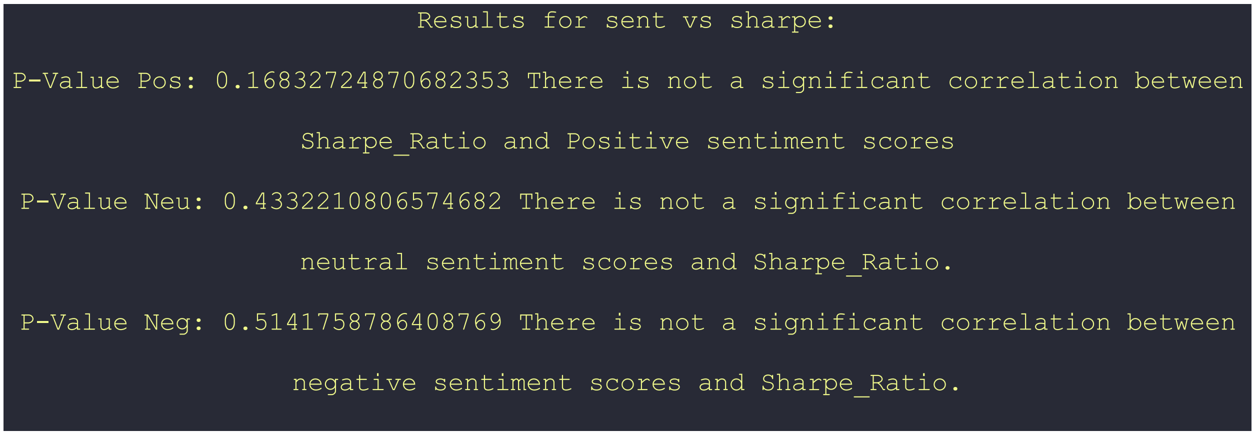




```
Results for sent vs returns:
P-Value Pos: 0.16832724870682353 There is not a significant correlation between
Portfolio_Return and Positive sentiment scores
P-Value Neu: 0.4332210806574682 There is not a significant correlation between
neutral sentiment scores and Portfolio_Return.
P-Value Neg: 0.5141758786408769 There is not a significant correlation between
negative sentiment scores and Portfolio_Return.
```

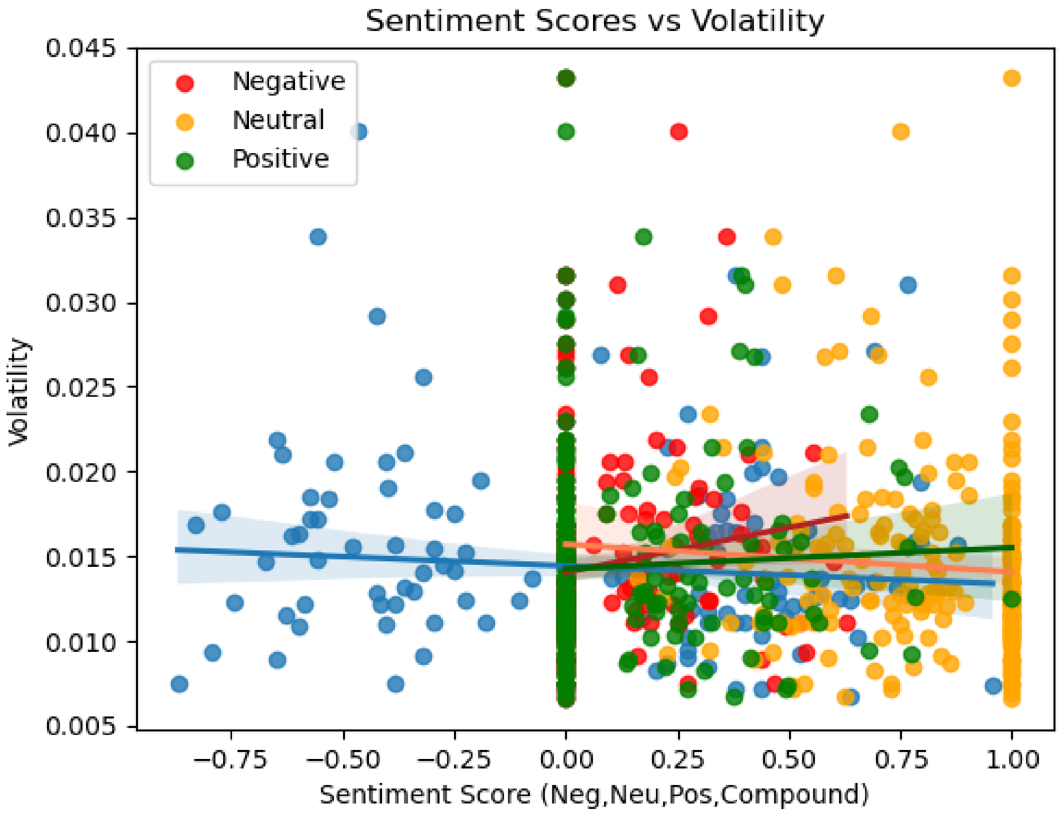
Sentiment vs Sharpe Ratio (risk adjusted version of sharpe ratio returns)

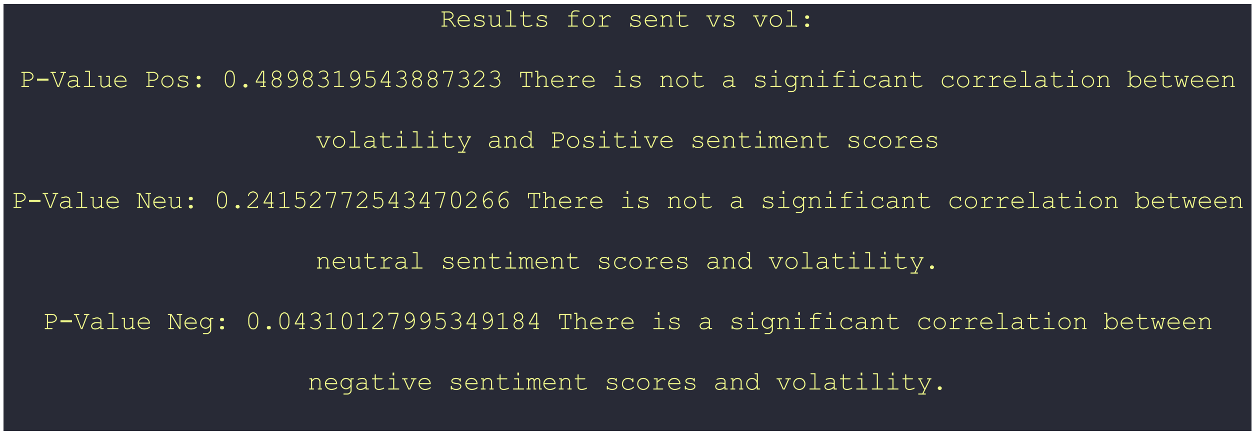




```
Results for sent vs sharpe:

P-Value Pos: 0.16832724870682353 There is not a significant correlation between

Sharpe_Ratio and Positive sentiment scores

P-Value Neu: 0.4332210806574682 There is not a significant correlation between

neutral sentiment scores and Sharpe_Ratio.

P-Value Neg: 0.5141758786408769 There is not a significant correlation between

negative sentiment scores and Sharpe_Ratio.
```

Sentiment vs Volatility





```
Results for sent vs vol:

P-Value Pos: 0.4898319543887323 There is not a significant correlation between

volatility and Positive sentiment scores

P-Value Neu: 0.24152772543470266 There is not a significant correlation between

neutral sentiment scores and volatility.

P-Value Neg: 0.04310127995349184 There is a significant correlation between

negative sentiment scores and volatility.
```

Correlation coefficient T-testing was used to compare two different groups to determine whether the difference between the means is statistically significant or not or had occurred by chance, and a significance value of 0.5 was used. Running this program revealed that there weren't any statistically significant correlations between expressed sentiment on Wall Street Bets and risk adjusted returns overall, as indicated the calculated p-values for the positive, neutral, and negative sentiment scores for portfolio returns and Sharpe ratios.

However there was a statistically significant increase in volatility when more negative sentiment is expressed in Wall Street Bets, as the p-value was calculated to be less than 0.5, indicating a potential impact of media sentiments on market trends.

Conclusions

The goals of this study were to investigate relationships between the sentiment found on Reddit's WallStreetBets subreddit and stock market risk adjusted returns. Much of the key findings obtained rejected the alternative hypothesis that a statistically significant correlation between Reddit sentiment scores and stock returns existed. But one of the findings that did accept the alternative hypothesis was a significant increase shown in volatility when there was more negative sentiments expressed in the WallStreetBets community.

A limitation faced in this project was with the way reddit is designed, their database does not allow us to search the posts by specific dates. [4] The approach to this project had to be modified to fetch only the top posts within the year using the praw package and we collected to provide insight on the sentiment dynamics with a more recent timeframe.

Another limitation that we faced is that there is no definitive answer to whether the changes in the historical price of our portfolio was a result of reddit sentiment or if reddit sentiment was the leading contributor to market shifts. Our main focus was finding a correlation rather than determining causation. Also, the visualizations were cluttered and are a bit hard to understand because of all the information displayed for sentiment score. To improve our analysis we could've compared portfolio returns, sharpe ratio, and volatility to each individual sentiment, positive, negative, or neutral instead of consolidating all of our data into one singular graph.

We can apply our data analysis to conclude when there is an increase in volatility due to the fact that increased volatility is important in the world of financial derivatives, specifically in trading of options. Implied Volatility is an important factor to remain profitable in this type of trading as it reflects the markets expectations and perception of future price fluctuations. Furthermore, it can be used to measure the expected magnitude of price swings leaving a greater margin for profit. A possible strategy utilizing our reddit volatility indicator could be to trade the opposite leg of a spread in order to capitalize on the increased levels of volatility within the market.

Author Contributions

As a group of three each person took on particular responsibilities to bring this project into completion. As the proposer of the topic, Nathaniel took a lead in this project and presented the initial data we were to work with. Nathaniel and Calvin collaborated on the development and execution of the program, and Rishpiath took a role of synthesizing outcomes and contributing to the insights reported.

References

1. Briggs, J. (2021, February 12). How-to use the reddit API in Python. YouTube. http://www.youtube.com/watch?v=FdjVoOf9HN4%2B%E2%80%8C

2. Beri, A. (2020a, May 27). Sentimental analysis using vader. Medium. https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

3. NLTK. (n.d.). https://www.nltk.org/

4. Get reddit posts in a given date range using python. Stack Overflow. https://stackoverflow.com/questions/61279115/get-reddit-posts-in-a-given-date-range-using-python
5. Bloomberg. (n.d.). United States rates &amp; bonds. Bloomberg.com. https://www.bloomberg.com/markets/rates-bonds/government-bonds/us

6. Fernando, J. (2023, May 11). Sharpe ratio formula and definition with examples. Investopedia. https://www.investopedia.com/terms/s/sharperatio.asp

7. Yahoo! (n.d.). Stocks in the Berkshire Hathaway portfolio. Yahoo! Finance. https://finance.yahoo.com/u/yahoo-finance/watchlists/the-berkshire-hathaway-portfolio/

8. API documentation¶. API Documentation - yahoofinance documentation. (n.d.). https://python-yahoofinance.readthedocs.io/en/latest/api.html#historical-data

9. Merge two data frames based on common column values in Pandas. Stack Overflow. https://stackoverflow.com/questions/43297589/merge-two-data-frames-based-on-common-column-values-in-pandas

10. Team, T. I. (2022, November 22). *How to calculate expected portfolio return*. Investopedia. https://www.investopedia.com/ask/answers/061215/how-can-i-calculate-expected-return-my-portfolio.asp