r/WallStreetBets Post Sentiment Analysis vs. Stock Market Performance

Cristina Lawson
Computer Science and Engineering
University of California, Riverside
Riverside, California, United States
claws004@ucr.edu

Jerry Tan
Computer Science and Engineering
University of California, Riverside
Riverside, California, United States
jtan027@ucr.edu

Abstract

In this paper, we will discuss our application of sentiment analysis and the machine learning techniques of cluster analysis (k-mean) and linear regression analysis in finding the correlation between r/WallStreetBets and the stock market. We used r/WallStreetBets data in order to see sentiment on specific stock tickers and see if it correlated to the movement and/or volatility of that stock.

Introduction

For our project, we implemented data mining techniques in order to see if and/or how the sentiment analysis of the stock market Reddit discussion forum, r/WallStreetBets, can predict the stock market and/or if they play a role in the stock market changes and volatility. This will allow us to see if an internet discussion forum can affect the stock market and if so by how much. We contributed two types of analyses to compare r/WallStreetBets sentiment analysis and stock market changes and volatility. These analyses include a cluster analysis (k-means) and a linear regression analysis using the implemented r/WSB sentiment analysis and the ticker prices.

Proposed Method

The methods we used to solve the problem were to implement two machine learning algorithms in order to observe the stock market prices and volatility given the sentiment of the posts on r/WSB.

Datasets

For this project, we used three datasets which included a dataset containing posts from r/WallStreetBets [1], a dataset of stock prices of S&P 500 companies [2], and a dataset of stock prices of GME [3].

Sentiment Analysis

In order to implement the sentiment analysis to be used in for our analysis using the machine learning algorithms, we used a library called Afinn and Vader in order to give us the sentiment scores as well as the positivity, neutrality, negativity, and compound of the posts associated with specific ticker(s).

For the reading in and formatting the data for the sentiment analysis, we first read in the names of the ticker and then the r/WSB post and data. We then made a list of the titles and removed emojis in the process to make string data analysis easier. These titles were then compared with the tickers to see if there were any ticker symbols within the symbols. If there were any

matches then those tickers were added to a list. The same method as the title was done with the body of the post.

For the sentiment analysis scores, we used the Afinn and Vader imported libraries, as stated earlier. The titles and the bodies were read into the libraries in order to get the outputted sentiment analysis. The sentimental analysis of the post was then combined in a data frame with the corresponding post data as well as the ticker symbols contained within that post.

The data included in the sentiment analysis data frame for the title of the post includes the title of the post, the tickers contained within the title of the post, the title sentiment score, and the title negative, neutral, positive, and compound scores.

The data included in the sentiment analysis data frame for the body of the post includes the body of the post, the tickers contained within the body of the post, the body sentiment score, and the body negative, neutral, positive, and compound scores

All of the data stated above was then combined into a single dataframe which was used for the data analysis for our project.

Title S&P 500 + GME Tickers	Title Sentiment Score	Title Negative Score	Title Neutral Score	Title Positive Score	Title Compound Score
{'GME'}	0.0	0.0	1.0	0.0	0.0
{'GME'}	0.0	0.0	1.0	0.0	0.0
{'GME'}	2.0	0.0	0.865	0.135	0.3291
{'GME'}	2.0	0.0	0.706	0.294	0.3612
[]	-7.0	0.399	0.497	0.104	-0.7983

Body S&P 500 + GME Tickers	Body Sentiment	Body Negative Score	Body Neutral Score	Body Positive Score	Body Compound Score
{'GME', 'JPM'}	5.0	0.06	0.811	0.129	0.9632
{'GME'}	0.0	0.075	0.857	0.068	-0.3919
{'GME', 'T'}	-23.0	0.167	0.776	0.057	-0.9777
{'GME'}	-88.0	0.137	0.733	0.13	-0.9462
{'GME', 'TSLA'}	-86.0	0.248	0.658	0.094	-0.9983

Figure 1. Samples of collected r/WSB posts and their Afinn and Vader scores

Cluster Analysis (K-Means)

Once sentiment analysis was complete, we implemented a Cluster Analysis (K-Means) model to see if there was a correlation between the sentiment of the posts on r/WSB and stock performance. For my models I just used the GME stock for my analysis since a majority of the discussion that took place on r/WSB centered around GME.

The Cluster Analysis models I performed was sentiment score vs. volume and sentiment score vs. price differences.

We first perform data analysis work by merging financial data with sentiment analysis data. Then we set input and output variables, so we analyze sentiment scores vs various stock performance metrics(separately).

The prediction was then output to plots.

Linear Regression Analysis

Once sentiment analysis is complete, then we can construct Linear Regression models to see if there is a linear correlation between stock sentiment from Reddit and stock performance from S&P 500. We were able to perform simple linear regression models that compare the

relationship between the final sentiment compound score vs various stock performance metrics like price/earnings or earnings per share etc.

We first perform data analysis work by merging financial data with sentiment analysis data. Then we set input and output variables, so we analyze sentiment scores vs various stock performance metrics(separately).

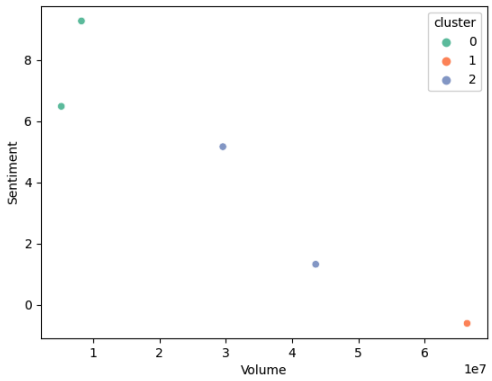
In order to build these models, we have to figure out the optimal theta values that we can use along with our x values to best accurately predict our y values/stock metric value. We implement a Gradient Descent algorithm that will run for some number of iterations in order to generate the most optimal theta value. We can then multiply the theta value by our x values to make our predictions. We can confirm an optimal value whenever we can minimize the sum of squared errors of our model.

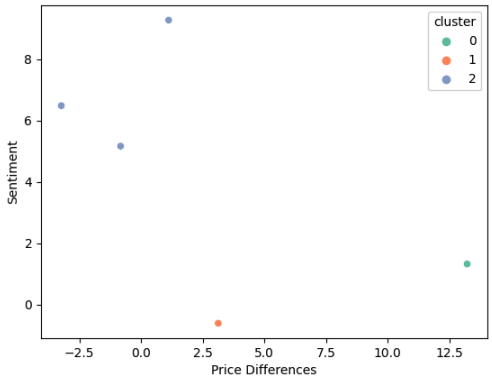
Our prediction can be mapped as a linear line, so we plot data points and our prediction line in 2d plots using matplotlib. By examining how well our line fits in with our data points or by examining the minimal cost of running gradient descent, we can determine how well stock sentiment impacts stock performance. All models are mapped one to one.

Experimental Evaluation

Cluster Analysis (K-Means) Analysis

The Cluster Analysis (K-Means) plots:





There seems to be a correlation between a lower sentiment and the volume bought being higher as well as the pierce difference being larger.

Something that was revealed within the data was that there is a slight correlation between the sentiment of r/WSB posts and stock market performance. There is a slight edge in the market, however, when the r/WSB posts are negative. This slight edge being that the stock is slightly more likely to have an up day when the r/WSB posts are negative. I conclude from this that the stock has been down for a couple of days due to a couple days of bad the r/WSB posts thus leading to a "discounted stock" that is more appealing to buy to investors. This increase in the amount of stocks bought on that day would thus lead to an increase in the price of the stock. As for positive the r/WSB posts, investors usually sell their stocks at a high, or

when they think the stock is above its value. This then leads to a drop in the price of the stock due to a high frequency of sells. All in all, I believe that the r/WSB posts slightly affect the stock prices, but not that much. Many of the stock ups and downs are due to a majority of people just wanting to buy or sell based on the price. The stock market is just based on the amount of stocks bought and sold, and not of the r/WSB posts so it is logical to deduce that the r/WSB posts have some effect.

Linear Regression Analysis

The main evaluation metric for the linear regression models is calculating the sum of the squared errors for all the linear regression samples. It is an accumulative evaluation metric that measures how far the linear regression prediction data points are from the actual data points from our dataset.

Here are the corresponding sum of squared errors for all the linear regression samples with :

Body Sentiment vs EBITDA : 1.493909668204751e+24

Body Sentiment vs Earnings/Share: 333244

Body Sentiment vs Price/Earnings: 43564606

Body Sentiment vs Price/Sales: 184525

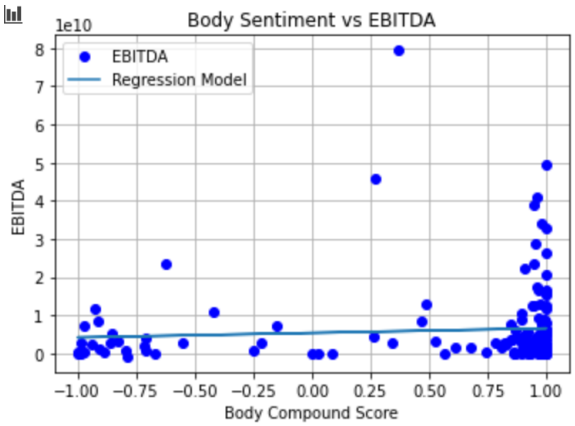
Title Sentiment vs EBITDA: 5.9891138676278905e+23

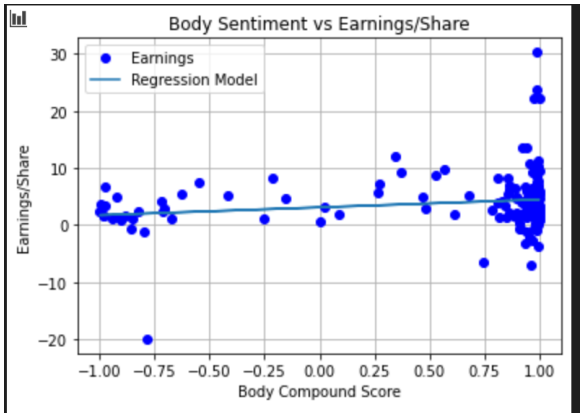
Title Sentiment vs Earnings/Share: 71123

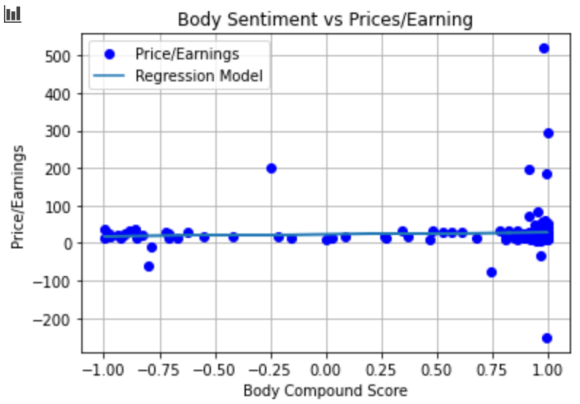
Title Sentiment vs Price/Earnings: 17820868

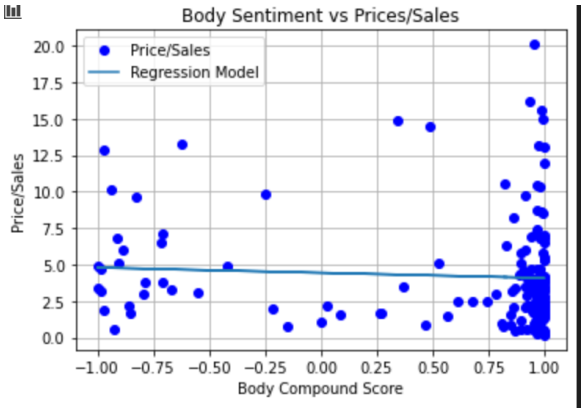
Title Sentiment vs Price/Sales: 49169

It seems as if for both title and body sentiment of stocks, sentiment and the price per sales of corresponding stock seem to be the most linear out of all the samples conducted. Overall, no linear regression examples seem to indicate a strong correlation between stock sentiment from Reddit and stock performance of corresponding stocks.









Related Work

There have been multiple other machine learning projects implemented that involve sentiment analysis of various textual sources and its effect on the stock market.

One topic that has been researched is related to the correlation between the news and the stock market. Two papers go over their research on predicting the effects of the stock market using news sentiment. Those two papers are “Stock Trend Prediction Using News Sentiment Analysis” [4] and “Predicting the Effects of News Sentiments on the Stock Market” [5]. Both of these papers do a sentiment analysis on current news and see if there is a correlation between news sentiment and the stock market.

Another topic that has been researched is related to the correlation between Twitter and the stock market. Two papers go over their research on predicting the effects of the effects of the stock market using Twitter tweet sentiment. Those papers are “Stock Prediction Using Twitter Sentiment Analysis” [6] and “Stock market prediction using Twitter sentiment analysis” [7]. Both of these papers do a sentiment analysis on tweets on Twitter and see if there is a correlation between tweet sentiment and the stock market.

All four of these research papers are related to our research paper since they all relate to sentiment analysis of real time information and see if they correlate with the stock market. They all want to see if either this public information affects the stock market in some way or they are just somehow correlated.

Discussions and Conclusions

Based on this project, we can conclude that the reddit WSB posts only slightly affect the performance and prices of stocks. As we were implementing the data mining techniques for this project, a few interesting thoughts occurred that would have made the project potentially more interesting. One reason why the linear regression models seem to not have strong correlation is possibly because the S&P financial data seems to not be the most up to date with the reddit posts. We could have done some web scraping to obtain the most up to date financial data in relation to the reddit posts. Another interesting thought that could have been implemented if time permitted was to perform multiple linear regression to check if various stock metrics could somehow affect sentiment of a stock.

Something that was revealed within the data was that there is a slight correlation between the sentiment of r/WSB posts and stock market performance. There is a slight edge in the

market, however, when the r/WSB posts are negative. This slight edge being that the stock is slightly more likely to have an up day when the r/WSB posts are negative. I conclude from this that the stock has been down for a couple of days due to a couple days of bad the r/WSB posts thus leading to a "discounted stock" that is more appealing to buy to investors. This increase in the amount of stocks bought on that day would thus lead to an increase in the price of the stock. As for positive the r/WSB posts, investors usually sell their stocks at a high, or when they think the stock is above its value. This then leads to a drop in the price of the stock due to a high frequency of sells. All in all, I believe that the r/WSB posts slightly affect the stock prices, but not that much. Many of the stock ups and downs are due to a majority of people just wanting to buy or sell based on the price. The stock market is just based on the amount of stocks bought and sold, and not of the r/WSB posts so it is logical to deduce that the r/WSB posts have some effect.

References

[1] G. Preeda. 2021. Reddit WallStreetBets Posts, *Kaggle*. https://www.kaggle.com/gpreda/reddit-wallstreetsbets-posts

[2] DataHub.io. 2021. S&P 500 Companies with Financial Information, *DataHub.io*. https://datahub.io/core/s-and-p-500-companies-financials

[3] Nasdaq. 2021. GME Historical Data, *Nasdaq*. https://www.nasdaq.com/market-activity/stocks/gme/historical

[4] K. Joshi, Prof. B. H. N., and Prof. J. Rao. "Stock Trend Prediction Using News Sentiment Analysis," KJSCE, Mumbai. https://arxiv.org/pdf/1607.01958.pdf

[5] D. Shah, H. Isah, and F. Zulkernine. "Predicting the Effects of News Sentiments on the Stock Market," Queens University. https://arxiv.org/pdf/1812.04199.pdf

[6] A. Mittal and A. Goel. "Stock Prediction Using Twitter Sentiment Analysis," Stanford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittalStockMarketPredictionUsingTwitterSentimentAnalysis. pdf), vol. 15, 2012.

[7] A. Kirlić, Z. Orhan, A. Hasovic, and M. Kevser-Gokgol. "Stock market prediction using Twitter sentiment analysis," IJRTEM. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3266569

[8] Shanmugam, Arun Ramji. "Simple Linear Regression with Example Using NumPy." *Medium*, Analytics Vidhya, 30 Aug. 2020, medium.com/analytics-vidhya/simple-linear-regression-with-example-using-numpy-e7b984f0d15e.