Using Sentiment Analysis on Reddit's r/wallstreetbets to Predict Changes in Stock Pricing.

Kyle Grace
kgrace@g.hmc.edu

Abstract

Through the use of the VADER sentiment analysis tool, specifically trained for social media data, we seek to uncover whether or not Redditors can predict the market and if this information can be used to generate a return on swing trades. Recently there's been a large increase in users sharing their ideas about investing on Reddit. Using sentiment analysis tools, we hope to discover whether or not data from Reddit can be used to predict the changes in individual stocks and the market as a whole.

1 Introduction

The stocks analyzed in this paper are traded on the New York Stock Exchange (NYSE) which is open for trading, generally, Monday through Friday from 9:30 a.m. to 4:00 p.m. Eastern Time. It is also closed on some holidays. Stocks rise and fall as entities buy and sell them, respectively. Because of this push and pull through market forces, many people have tried to use social media and the internet to predict stock price changes ahead of time. In this project, we focus on short-term trades, buying stocks and selling within a week or so. This is sometimes called swing trading. The stock market is generally believed to be mostly efficient, with most publicly available information already accounted for in the stock price, meaning individuals shouldn't be able to make profit off of information that is publicly available, they would just be speculating. In general, stocks have generally trended upwards over time, but with relatively unpredictable dips due to sell-offs or events affecting the company and larger economy.

Reddit is a website that hosts a collection of message boards, called Subreddits, where users can share text, images, and other content. r/wallstreetbets is a board focused on users posting their wild stock predictions and options trading activity, a stock trading technique designed to be very high-risk high-reward. By understanding the sentiment of Reddit users we hope to gain an idea of sentiment of individuals interested in stock trading.

For our analysis, we have chosen 7 stocks which have been relatively popular over the last 5 years. Those stocks are SPY (SP 500 Index), SBUX (Starbucks), GOOGL (Alphabet, the holding company of Google), TSLA (Tesla), AAPL (Apple), NFLX (Netflix), FB (Facebook). The abbreviation of a stock is called a ticker and is used to represent the stocks on the stock exchange and online.

2 Related Work

There has been some previous work in similar areas to this project, largely using things like Twitter data. (Mittal and Goel, 2012) and (Nguyen and Shirai, 2015) are papers that look at the entirety of Twitter sentiment to predict market movements, focusing on the Dow Jones Industrial Average index. (Pagolu et al., 2016) focused on specific stocks and the Tweets being made that mention them, which reflects the approach in this paper of looking at specific stocks. (Li et al., 2014) looks at the sentiment of news to predict stock pricing, which is similar but does not rely on social media or internet message boards.

There are also some non-academic efforts to solve this problem, like that of Redditor u/swaggymedia who created swaggystocks.com which is a realtime dashboard tracking various stock tickers along with their sentiment from recent posts on r/wallstreetbets. They use a bag of words sentiment analysis but did not publish more specifics.

More generally, there are existing classifiers that have been trained on social media data that will be useful in the context of Reddit, which would more closely resemble other social media than traditional literature. The VADER algorithm is one

that has shown to be very effective on social media text (Gilbert and Hutto, 2014). (Elbagir and Yang, 2019) uses the VADER algorithm to analyze sentiment of Tweets related to the 2016 Presidential Election in the United States. This analysis was taken even further in (Yaqub et al., 2020) to run location-based sentiment analysis in 10 states to try and predict the winner in the 2016 election in that state, correctly predicting 8 of the 10 states in question.

3 Methods

3.1 Data

To complete this project, data from the Subreddit r/wallstreetbets is used. This Subreddit has had over 450,000 posts since it was created, the data has been compiled in a Kaggle database created by Sheridan Green. [1] Each post has other data associated with it such as the date and time of posting as well as community "upvotes and downvotes" which is a way for the community to vote to for posts.

Historical data for the stocks has been pulled Yahoo Finance which is able to provide daily, weekly, and monthly data on nearly all publicly available stocks going back as far as they've been available.

3.2 Tools

To complete the sentiment analysis, the VADER algorithm was used. This will be used to generate sentiment scores for individual posts. The model was trained on social media data so it is hopefully well-suited to handle the informal nature of our data.

3.3 Procedure

To complete the analysis, the following process was followed.

1. Iterate through each entry on the subreddit.

2. Each post was associated with a day. Weekends and holidays were pushed to being associated with the next trading day, since posts can be made at any time but the market only moves during trading hours on weekdays.

3. The VADER python package was used to analyze the sentiment of each post's content along with its title. The compound score from the algorithm was then multiplied by the logarithm of the up-votes as a means to weight scores by their popularity. The weighted score, is given by the following equation.

$$(S(post) + S(title)) \times log_{10}(upvotes)$$

Where $S(x)$ is the polarity score from VADER for a text input x. The log of up-votes was used to not overstate viral posts.

The post was then scrapped to find our tickers of interest within the text, we then added the weighted score of the post to a daily aggregate for the mentioned stock. All posts were also aggregated into an "overall" sentiment score that took into account every post, regardless of if it was one of our stocks of interest.

4. Initially we took a binary approach to analysis, we noted the percentage of times where the sign of our sentiment (positive or negative) lead to a correspondingly signed return.

5. The initiative data were then used as inputs to a regression on stock returns over daily and weekly future time horizons. A given day's sentiment score would be regressed on both the following day and following 5 days return on that stock. These are done using an ordinary-least-squares method at the 95% confidence interval. This will tell us how well the variance in sentiment data reflects the variance in stock prices.

Ticker Symbol	Weighted Sentiment Score
SPY	0.33
SBUX	0.14
GOOGL	0.95
TSLA	0.51
AAPL	0.68
NFLX	0.25
FB	0.66
All Posts	18.12

Table 1: Average Daily Sentiment

4 Results

Table 1 shows the average weighted sentiment score for each of our tickers. We can see that on

[1] https://www.kaggle.com/shergreen/wallstreetbets-subreddit-submissions

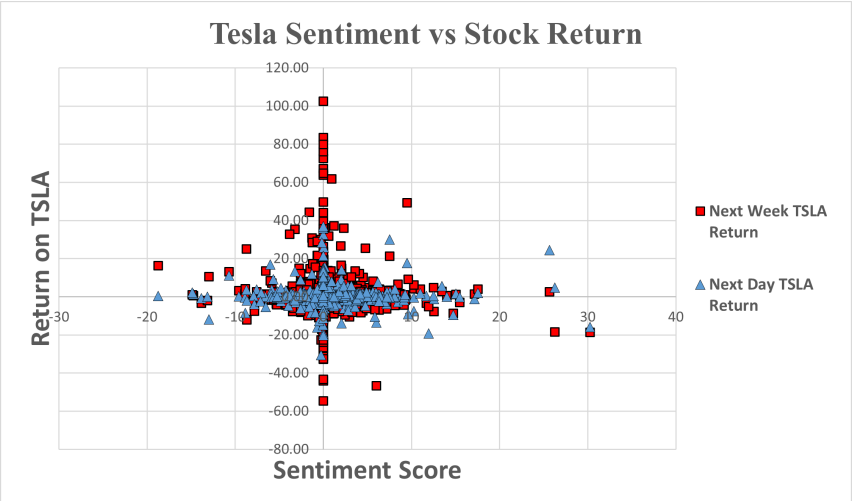

Figure 1: Daily sentiment score vs the upcoming day and week return on TSLA from 2015 to 2020

Ticker Symbol	Corrent Binary Predictions
SPY	53.6%
SBUX	53.0%
GOOGL	51.8%
TSLA	51.8%
AAPL	52.2%
NFLX	52.6%
FB	53.3%
All Posts vs. SPY	54.2%

Table 2: Binary Sentiment Daily Prediction Results. Comparing the sign of the sentiment with the sign of the following day's returns.

average, every ticker had a positive sentiment associated with it and the average weighted sentiment on the Subreddit was positive.

The binary results of our data show that just slightly greater than 50% of the time, the sign of sentiment matched the sign of the following day's return. However, this could be due to our data's overall skew towards positive sentiment and also stocks generally having a positive return over the last 5 years. Table 2 details the breakdown of accuracy given the individual tickers as well as overall sentiment of all posts compared to SPY.

Figure 1 shows the overall scatter of daily sentiment scores compared to the following day and 5 days returns. We can see that overall the data is clustered around the center, skewed towards an overall positive sentiment. There is no clear overall trend visible in the data.

Figure 2 focuses in on one of the stocks in specific, TSLA, and a similar lack of trend is present. There is a near uniform distribution of positive and negative sentiment with an overall bias towards positive returns. Table 3 clearly shows that when it comes to using sentiment scores in a regression, there is little to no correlation. Consistently high p-values indicate that we fail to reject the null hypothesis on all stocks, except for TSLA. TSLA is the only one with p-value under 0.05 so we fail to reject the null hypothesis, yet we still have a very low R^2.

We used our regression as a predictive model for TSLA, since we cannot reject the null hypothesis. The intercept was 1.43 and our coefficient is -0.1927. In a binary prediction, it is correct 56.2% of the time, an improvement of about 5% over pure sentiment alone. If a hypothetical investor used this model to invest over the past 5 years, investing our estimated return in dollars, then selling 5 days later. They would've made $1746.10 from a cumulative investment of $1723.18. This is a 0.2% annual return, which is lower than the average return on government bonds over that time, considered a risk-less investment.

Figure 3 is another visualization of the data for the FB ticker. It is clear both returns and sentiment have much volatility with many of the peaks not lining up or being in the opposite direction.

4.1 Discussion

The data shows that VADER sentiment scores do not provide generally useful predictions on the changes in stock prices. It is slightly better than random guessing in most cases, but not enough to beat a market index or government bond investing. This follows with the market efficiency theory that

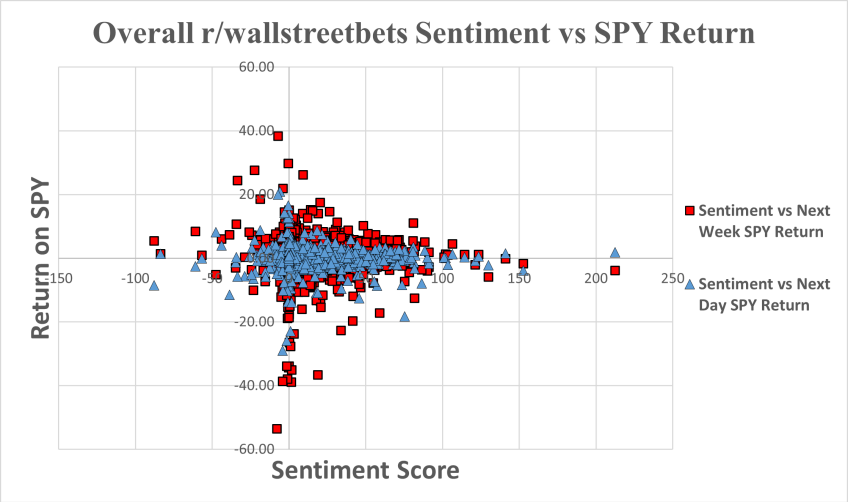

Figure 2: Daily sentiment score vs the upcoming day and week return on SPY from 2015 to 2020

Ticker Symbol	p-value	R^2	Coefficient
SPY	0.270	0.031	-0.054
SBUX	0.083	0.00233	-0.0875
GOOGL	0.467	0.0203	-0.257
TSLA	0.027	0.00379	-0.1927
AAPL	0.586	0.015	0.012
NFLX	0.330	0.027	-0.266
FB	0.879	0.004	0.011
All Posts vs. SPY	0.374	0.00061	-0.0063

Table 3: Results of the returns on the stock over the next 5 days on our daily sentiment score from 2015 to 2020.

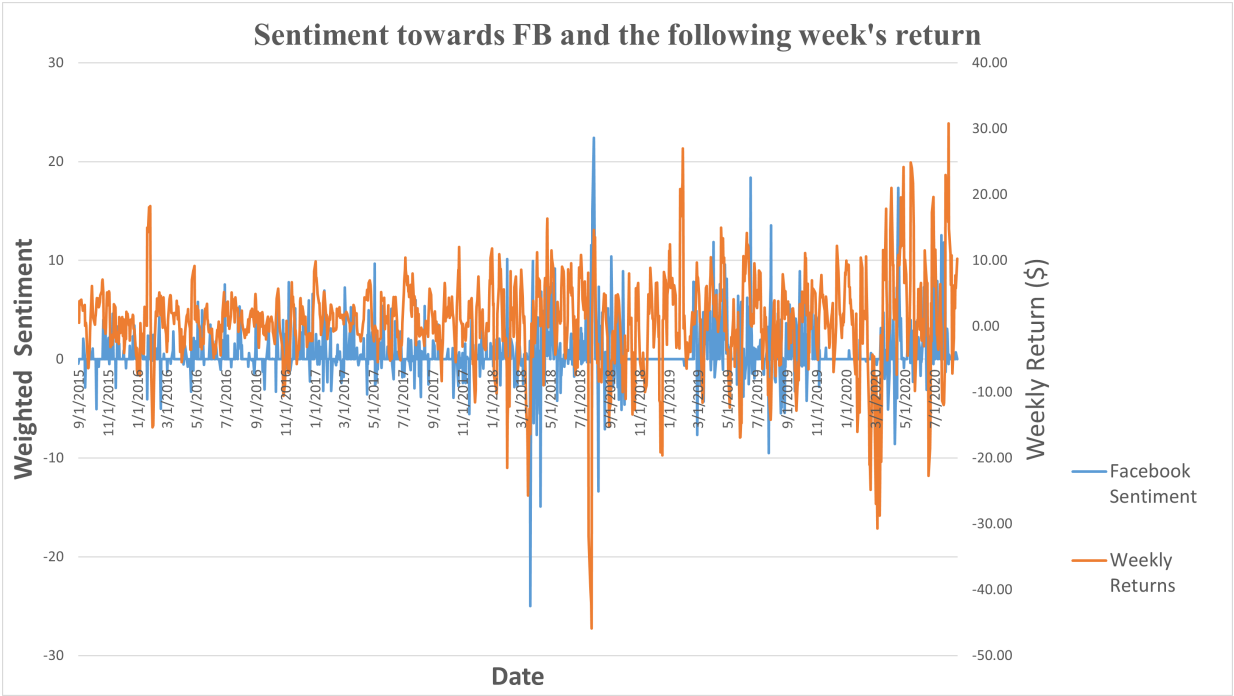

Figure 3: Daily Sentiment and the following 5 day return over time.

returns cannot be generated by using publicly available information, it should already be incorporated in the pricing of the stock. We fail to reject the null hypothesis that the sentiment of Reddit posts has no relation to future stock returns in the next day or week. Our results are lower than the 56% accuracy achieved by (Nguyen and Shirai, 2015). The only promising result is the near 53% accuracy achieved in using sentiment to predict the simple question of if stocks will move up or down, however this is hardly novel given overall upward movement in stock prices and the overall positive average sentiment of the posts.

4.2 Future Work

Future extensions of this project could include:

- A live tracking tool similar to swaggystocks.com which utilizes the VADER sentiment analysis tool, could be used to view how well it tracks in real time. If the model was able to be modified to better fit the data, then a tool like this could be useful for making trades.
- VADER is trained on Twitter data but r/wallstreebets tend to have lots of jargon so there could be improved accuracy to be had in classifying the text by including handling for specific jargon.
- This data could also be combined with other indicators into a multidimensional model to help augment the ability of other variables.

References

Shihab Elbagir and Jing Yang. 2019. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, pages 122–16.

CHE Gilbert and Erric Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, volume 81, page 82.

Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang, and Xiaotie Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Anshul Mittal and Arpit Goel. 2012. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE.

Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. 2020. Location-based sentiment analyses and visualization of twitter election data. *Digital Government: Research and Practice*, 1(2):1–19.