Reddit Social Media Scraper

Devon Johnson, Kai Moy, Tyson Mehrens, Zhenyu Zhen

June 13, 2021

Executive Summary:

Our project group set out to discover if there was a correlation between the sentiment and frequency of reddit users posts on a subreddit called "WallStreetBets'' and the stock price of GameStop. To test this theory, we used a sample size of around 7000 comments during a 60 day period from December 15th, 2020 to February 15th, 2021 when GameStop's stock price made a nearly 1000% rise. Each comment was inserted into a VADER sentiment analyzer in order to find a binary sentiment(positive or negative) for each comment. Next, we developed a linear regression model to see if we could correctly predict the change in price for a given day based on the four variables: number of positive comments, number of negative comments, overall positive sentiment average and overall negative sentiment average for a given day. To test this model we compared our predicted change in price with the real daily price for each of the 60 days during our period. After training this model, we ran a number of statistical tests in order to determine how accurately our variables predicted the real change in price. We concluded that our linear regression model did not have the statistical significance necessary to prove that it can accurately predict the change in price of GameStop. Intuitively however, it was clear that the Reddit page "WallStreetBets" did in fact play a significant role in GameStop's dramatic price volatility in early 2021. Whoever can build a model that is statistically significant enough to accurately predict the change in price based on social media sentiment will be able to make a fortune.

Member Listing

Member Name	**Member Email**	**Member Role**
Devon Johnson	djohnson@calpoly.edu	Writer/Coder
Kai Moy	kamoy@calpoly.edu	Coder
Tyson Mehrens	tmehrens@calpoly.edu	Coder
Zhenyu Zhen	zzhen01@calpoly.edu	Coder

Problem

During the first quarter of 2021, the stock price of GameStop, a publicly traded stock, peaked at a record high of $347.51 on January 27, 2021. This remarkable and unprecedented turn of events led to shocking effects in the market, some being slightly problematic or astronomically beneficial. The stock phenomenon was a Robinhood scheme to take from the rich and give to the poor, the rich being large hedge funds and the poor being reddit users. Large hedge funds tried to short or bet against Gamestop stocks. Shorting a stock essentially means borrowing shares from a broker and selling them, agreeing that the shares will be returned at a later date. If the price falls, the entity who has shorted the stock, buys back the shares and pockets the difference. But shorting a stock has its risks, if the price rises, losses are catastrophic. Amateur investors out maneuvered large hedge funds by pouring investments into the shorted stock, costing them millions to billions of dollars. Amateur investors using the subreddit "wallstreetbets" began buying GameStop stocks, raising the price and placing bets on the opposite side of the shorts.

They succeeded in their efforts against the large hedge funds and we suspect that social media was a large contributor to their success. The question which we will be answering is how social media allowed enough people to join the cause or "ride the wave" to make a dramatic difference in the price of GameStop stock. Melvin Capital, one of the Hedge funds shorting GameStop, needed a 2.75 billion cash injection to continue business operations. For the hedge funds, this was a billion dollar problem and exposed the flaws in the stock market. This flaw and others of similar essence were to be used as observation in our research. The way that our project will seek to answer the question is by evaluating the frequency and sentiment(positive or negative) of social media posts on reddit and their impact on the price of GameStop.

We expect to find a high correlation between the volume and ratio of positive and negative comments involving GME and the price of the stock. The potential significance of this is that the patterns identified in the price of GME relative to social media sentiment can be used to predict the performance of future "meme" stocks/investments. Figuring out an algorithm to predict future market activity has the potential to transform trading on a tremendous scale.

Our datasource will come from a subreddit called WallStreetBets. WallStreetBets is a free flowing forum where retail investors discuss their

investments in stocks and other assets. We have chosen this forum because this is where the excitement for GME began and served as a rallying ground for retail investors to discuss their thoughts and enthusiasm for the stock.

For each comment we will seek the timestamp of when it was posted. We will also seek to determine the sentiment of each data record by using text mining processes. We want to determine if the comment offers a "bullish" or "bearish" stance on the price of GME. A bullish stance means that the commenter believes the price will go up, while a bearish stance means that the commenter believes the price will go down. We can then aggregate a count of all bullish(positive) or bearish(negative) comments and determine a ratio of positive to negative comments. We then can compare that to GME's historical candlestick trading charts and capture each day as a datarecord. We will then compare the overall percentage increase or decrease in share price to the overall sentiment on that day from all comments to see if they are correlated and can be used to better predict a stock's price. We can also use the frequency of comments on each day during our period and compare to the percentage change on that day. We can compare the %change in price on high frequency days to lower frequency days. We hope to be able to calculate how many negative, or positive comments are required to move the stock price an entire percent.

Data Set

Our team's goal is to discover the sentiment of the 7,000 comments and determine how the number of positive and negative comments impacted GameStop's stock prices. We decided to use sentiment analysis to discover the number of positive and negative comments per day between December 15, 2020, and February 15, 2021, and use an analysis tool to discover the impacts these comments had on GameStop's stock prices. Since we are trying to determine the sentiment (positive, negative or neutral) of a large amount of text making sentiment analysis the optimal text mining tool. However, there are many different types of sentiment analysis. The type of sentiment analysis we plan to use is **VADER** (Valence Aware Dictionary and Sentiment Reasoner) to create a fine-grained polarity scale. VADER allowed us to sort the sentiment of our data more accurately and precisely. When we are working with such a large set and if we were sorting through it manually, we are at risk of missing key words, abbreviations, and phrases that would have more specifically depicted the sentiment of each individual reddit comments from the threads. Fortunately, VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It can handle hashtags, emojis, and intentional misspellings. Our dataset is an unstructured social media dataset that includes emojis like rocket ships, diamonds with a hand and suns, etc. Which means that VADER is able to process all of these, making it the correct choice. Vader eliminates this issue by allowing us to input our reddit and specify rules that allow us to generate our sentiment outputs. During this process, we scored our reddit comments based on their compound score. If the reddit thread comment's compound score was greater than one, we labeled it as "Positive". If a reddit thread comment's compound score was less than zero we labeled it as "Negative". Additionally, if a reddit thread comment's compound score was zero, we labeled it as

```
analyzer = SentimentIntensityAnalyzer()
vadercompound_list = []
for comment in comment_body_list:
        vs = analyzer.polarity_scores(comment)
        compound_score = vs["compound"]
        sentiment = ""
        if(compound_score > 0.05):
            sentiment = "Positive"
        elif(compound_score < -0.05):
            sentiment = "Negative"
        else:
            sentiment = "Neutral"
        vadercompound_list.append(compound_score)
```

"Neutral". Given what we know about Gamestop stock prices and the range of feelings towards it, and our results are around how we expected it to look like.

With this information, we made multiple lists: the number of positive comments per day, the number of negative comments per day, the number of neutral comments per day, the number of total comments per day, the average positive score per day, the average negative score per day, and the average neutral score per day. With these lists, we were able to formulate our linear analysis model.

```
Y = ground_truth
Y = np.asarray(Y)
#Constant Variable
x0 = [constant]
#Count of Positive Comments on a given day
x1 = [pos_comment]
#Count of Negative Comments
x2 = [neg_comment]
x3 = [daily_avg_pos]
x4 = [daily_avg_neg]
x0 = np.asarray(x0)
x1 = np.asarray(x1)
x2 = np.asarray(x2)
x3 = np.asarray(x3)
x4 = np.asarray(x4)
print(X)
```

We chose to conduct a **Linear Regression Analysis** because Regression analysis mathematically describes the relationship between independent variables and the dependent variable. It also allows us to predict the mean value of the dependent variable when you specify values for the independent variables. Our independent variable or Y was the average change in Gamestop price over the two-month span. Our linear equation, which includes all of our inputs(dependent variables), is Y(change in price)= x0(average in price)+ total positives comments* B1 + total negatives comments *B2+average score of positive sentiment*B3 +average score of negative sentiment *B4. We used the concatenate function to combine all of the x variables into a single X.

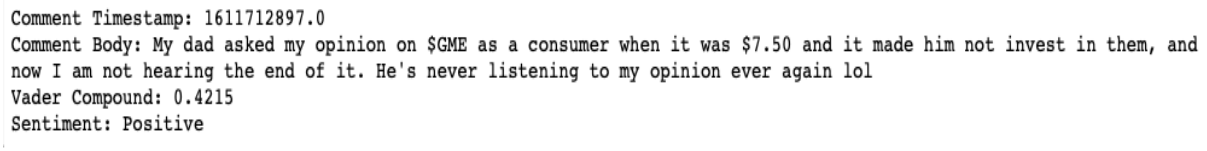
After we conducted a linear regression score, we used the number to conduct an **OLS Regression Results table**. With the table, we are able to see values such as the R-squared values , Adjusted R-squared values,

coefficient, stand error values and more values that show whether there is a relationship between social media posts and the stock price.

Results

The results from running Vader Sentiment Analysis on our data set were recorded into a list and depicted in Image 1. Referencing the Vader package, provided us with three classifications for sentiment: Positive, Negative, and Neutral which were weighed in regards to a 10% neutrality threshold. The compound Vader sentiment scores, or sum of the three scores for the comment bodies, were stored and used for calculating the variables of the Linear Regression Model. The quantitative results that we observed from the sentiment analysis varied in accuracy. Some of the predicted sentiments were correctly resembled by their vader compound score, and others completely wrong. We believe that this led to some of the lack of correlation which we observed in our Linear Regression Model.

Image 1-Vader Results



The results from our regression are depicted in Table 1. Our Regression Equations with Values: Y(change in price)= x0 + total positives comments* .0030 + total negatives comments *-.0035+average score of positive sentiment*-.1457+average score of negative sentiment *-.0850. Interpreting the betas is one of the most interesting parts of linear regression. Holding all other variables constant, an increase in one positive comment increases the change in GME stock price by .0030.
Holding all other variables constant, an increase in one negative comment decreases the change in GME stock price by .0035. Holding all other variables, a one-unit increase in the average positive sentiment score decreases the change in GameStop stock price by .1457. Holding all other variables, a one-unit increase in the average negative sentiment score decreases the change in GameStop stock price by .0850. Unfortunately, none of the variables are significant, p-value less than .05, but X2 is very close. The p-value for Total Positive Comments (X2) is .052. If we were to run another regression I would focus on how additional variables would make this coefficient more significant.

To analyze the model as a whole we will use the R-squared value. R squared is possibly the most important measurement produced by this summary. R-squared is the measurement of how much of the dependent variable (Y) is explained by changes in our independent variables (Xs), (“R-Squared vs. Adjusted R-Squared: What’s the Difference?”). As I stated

earlier, our independent variable is the change in stock price and our dependent variables are the number of positive comments, number of negative comments, average positive sentiment score per day, and average negative sentiment score per day. Adjusted R-squared is adjusted for the number of predictors in the model in our case 5. Our adjusted R squared of .104 means our model explains 10.4% of the change in the change of Gamestops stock price. This is a pretty low adjusted r-squared, which means our model has room for improvement. A good starting place would be determining how accurate the Vader analysis was. Another solution would be adding more variables like total number of comments per day and possibly separating negative and positive comments in 2 different regressions. The time a comment was published also seems important, because trading only occurs on weekdays from 9:30 am-4 pm.

In conclusion, this is a very interesting topic that haPeople spend their whole careers trying to figure out how social media affects the stock market, so this was a great starting place. We learned how to implement text mining in a meaningful way and are grateful for the opportunity.

Table 1- OLS Results

In [72]:
```
import statsmodels.api as sm
ols = sm.OLS(Y,X)
ols_result = ols.fit()
print(ols_result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.162
Model:                            OLS   Adj. R-squared:                  0.104
Method:                 Least Squares   F-statistic:                     2.796
Date:                Mon, 31 May 2021   Prob (F-statistic):             0.0342
Time:                        15:36:31   Log-Likelihood:                -2.1102
No. Observations:                  63   AIC:                             14.22
Df Residuals:                      58   BIC:                             24.94
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1632      1.082      0.151      0.881      -2.003       2.329
x2             0.0030      0.002      1.986      0.052   -2.39e-05       0.006
x3            -0.0035      0.003     -1.200      0.235      -0.009       0.002
x4            -0.1457      0.205     -0.711      0.480      -0.556       0.265
x5            -0.0850      0.233     -0.365      0.717      -0.551       0.381
==============================================================================
Omnibus:                       47.691   Durbin-Watson:                   1.762
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              263.755
Skew:                           1.988   Prob(JB):                     5.33e-58
Kurtosis:                      12.202   Cond. No.                     3.66e+03
==============================================================================
```

Works Cited

"R-Squared vs. Adjusted R-Squared: What's the Difference?" *Investopedia*, 2021, www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp. Accessed 11 June 2021.

https://praw.readthedocs.io/en/latest/

https://pypi.org/project/vaderSentiment/

https://www.statsmodels.org/stable/index.html