War on Wall Street:
Analysis of Today's Stock Market vs. Social Media Influence

George Padavick, Matt Gilgo
CS 506, Spring '21
Final Project Report

Table of Contents

Introduction

The year of 2020 was certainly an unprecedented time. A pandemic struck the masses which caused a paradigm shift in society's day-to-day interactions. Families were confined to their homes, left with copious amounts of time and in many cases, greater savings due to a lack of discretionary spending. This environment, paired with the widespread availability of free trading platforms like Robinhood and Webull, encouraged massive growth in small retail investors. However, the surge in inexperienced investors has resulted in untraditional market behavior that is often driven by market hype rather than stock fundamentals. This report aims to analyze some of the new market trends seen in 2020 and 2021, with a specific focus on the relationship between market behavior and social media.

A recent example of this unprecedented market behavior occurred in January 2021 with the pumping of then-dying stock Gamestop (GME) from $17.85 to $483.00 in less than a month. The growth in this stock was not based on improved revenue, a business model transition, or anything that would traditionally move the needle. Instead it was driven by a movement on various social media platforms. Users across the world on applications such as Twitter, Reddit, Facebook, etc. were seen tagging GME in posts and encouraging others to band together and get rich as they drove up the price to squeeze the "short" position hedge funds into bankruptcy. This mass craze garnered substantial media attention and left many traditional investors dumbfounded as to how everyday people could influence the market at scale simply by working together on social media.

However, many traditional investors and media outlets did not realize that this behavior had been occurring for many months prior at a smaller scale. Stocks in companies such as XspresSpa (XSPA), Nikola (NKLA), and Genius Brands (GNUS) had seen major price inflation due to a social media community driving in new investors through using similar rhetoric. Yet, similar to GME, these stocks also saw a similar price correction soon after the popularity had hit its peak. In only 3 weeks after posting its All-Time High, Gamestop had crashed back down to the $40 range, leaving investors at the top of the hype with only 10% of their original investment.

Based on the examples described above, it is clear that the post-pandemic market will include greater influence from social media platforms. The analysis outlined in this report attempts to better understand the relationship between social media and market performance, specifically in the case of Twitter data. This involves collecting tweets from known cases of "pump and dump" schemes to understand if the underlying Twitter sentiment is correlated with price movement.

Data Collection

Multiple social media platforms were considered for this analysis based on ease of data access and previous findings of social media influence based on personal experience. Of the platforms considered, Twitter was determined to be the best platform for this study due to the existence of an easy to use API along with past examples of market related activity. Specifically, Tweepy was used to collect tweets within the 7 day window allotted to the free tier of the Twitter API. Besides twitter data, financial data was also collected for comparison with twitter sentiment. While there is a breadth of financial APIs, research and testing was performed on APIs for Alpha Vantage, Finnhub, and YahooFinance to determine which API would be most applicable to this analysis. Ultimately, due to a high level of documentation, ease-of-use, and inclusion of key features needed for this analysis, YahooFinance's API was selected as the source of the financial data seen in this report. The overall selection process as well as additional API testing is outlined in the code/api_testing folder of the project GitHub repository.

Data Ingestion

While both Tweepy and Yahoo Finance provide many features to pull data, the data ingestion process for this analysis required additional steps to organize and clean the data. A high level illustration of this process is shown below in figure 1.

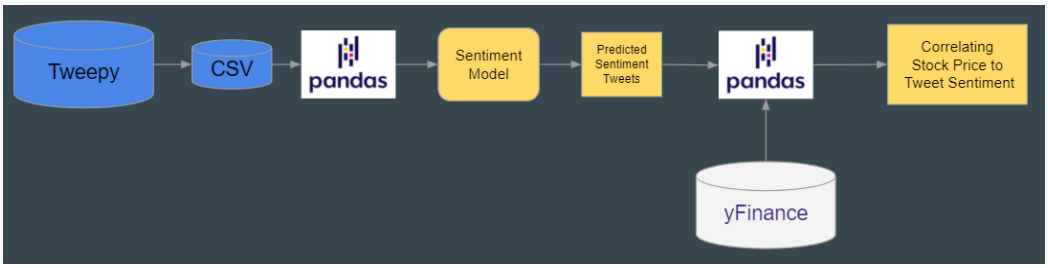

Figure 1: Data Ingestion Map

As shown above, Twitter data was extracted from the Tweepy API and stored in pickle files specific to each stock in consideration. These pickle files are available in the project repository under 'code/data_aquisition/twitter_data/clean_data' and can be imported into a single dataframe using the 'read_all_twitter_data' function in twitter_data.py. Due to Twitter API limitations that restrict data pulls to only tweets in the last seven days, only stocks that exhibited quick pump-and-dumps over the last two months were considered. Specifically, the following stocks were used to build a training set of known pump-and-dumps:

- EEENF
- GYST
- DLPN
- TTCM
- WNRS
- SEAC
- CERPQ

Data Preparation and Cleaning

While each tweet is bound to the 280 character limit imposed by Twitter, the variation in each tweet still makes it difficult to build features from the text. To address the variation in each tweet, a few preprocessing steps were taken to clean the data. First, all tweets were converted to lowercase and the following features were removed.

- Punctuation such as periods, commas, and ellipsis marks.
- Hyperlinks
- Stock tags such as 'GME
- Tags to other twitter users such as "@username"

Next, an additional column was created to list the date and time in eastern US time so that the twitter data could be easily merged and compared with market data. Finally, in order to extract information concerning emojis, the emojis in each tweet had to be decoded into their respective unicode names. For example, the 👍 was converted to ":thumbs_up:" where the ':' characters denote that the text contains an emoji. The process outlined above is summarized in the 'clean_twitter_text' function in 'features.py'.

Besides cleaning the tweet text, additional considerations were made with the metadata provided by twitter for each tweet. For example, tweets in response to another tweet such as retweets and quote tweets were included in the full dataset since they still offer important information regarding user sentiment towards a particular stock. Additionally, features like follower count and favourites count were included in the dataset as they offer insight into the popularity of a specific tweet.

Besides Twitter data, the Yahoo Finance data had a few decisions that needed to be made as well. For example, the stock data could be exported in various buckets of intervals (5-Minute, 30-Minute, 1-Day, etc.), so choosing a proper one to mate well with the twitter data was important. With both of these datasets in a Pandas dataframe, the merge_on function provided the capability to round tweets to their closest time bucket of the stock data. It was found that stocks with high tweet volume had great success being joined with 5-Minute interval data, while lower tweet volume stocks had more success with 1-Hour intervals.

Feature Extraction

To both better understand the data as well as improve modeling capability, a substantial amount of time was spent in the feature extraction phase. Specifically, features from the Twitter API along with features related to tweet sentiment were considered in this process.

Twitter Feature Review

Initially, various features from the Twitter API were considered to aid in the dowselection process. This involved plotting each feature against stock price and looking for correlations between price and twitter activity. An example using followers count is provided below.

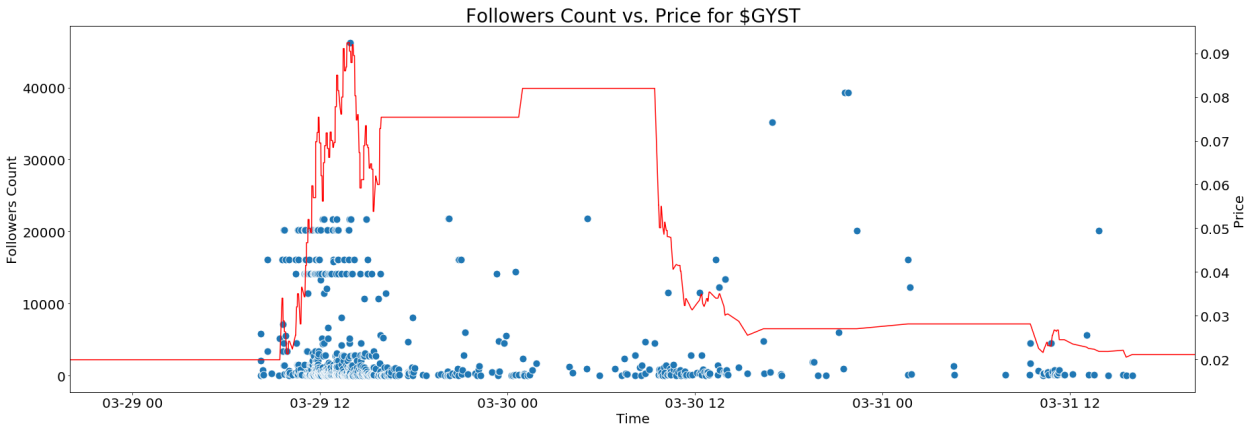

Figure 2: Feature vs Price Example

The plot above shows a scatter plot of tweets with the "$GYST" tag plotted against GYST price for the same date range. Specifically, this plot looks at the followers count for the user that generated each tweet which is shown on the left y-axis. This plot illustrates both the relationship between tweet volume and price as well as the distinction between tweets from accounts with larger followings and tweets from smaller accounts. For example, in the early stages of the pump, there is substantial tweet volume not only from smaller accounts but also from a few accounts within the 10,000 to 250,000 followers range. However, as the price drops the volume of tweets from users with higher followings also tends to decrease suggesting that followers count may be an important feature to train on. Besides followers count, favourites count and retweets count were also identified as good features to include in our models.

Sentiment Feature Engineering

After reviewing the available features in the twitter data, additional analysis was performed on the text in each tweet to better characterize the tweet sentiment. Initially, pretrained sentiment models were investigated along with sentiment training datasets available from resources like Kaggle. However, these models and datasets struggled to interpret twitter sentiment mainly due to the difference in language used on twitter versus everyday vocabulary. For example, a positive tweet will usually take the following form.

`big otc gainer! 🚀 🚀 🚀 $GYST`

The example above demonstrates both the limited vocabulary used as well as the importance of emojis in interpreting tweet sentiment. In this case the rocket emoji and “gainer” may be the best indicators that the tweet is positive. Unfortunately, many pretrained models and existing datasets do not capture this information in regards to sentiment analysis, especially in terms of the relationship between emojis and sentiment. For this reason, a training dataset was manually created using the tweets pulled from cases of known pump-and-dumps.

This process involved scoring over 12000 tweets for sentiment as well as other features like tags for known pumpers, price regions, and inflections points. Sentiment Score and Price Region were both rated on a -1, 0, 1 (Negative, Neutral, Positive) scale, with the intent of providing features to train both the sentiment of the tweets and points in time where it was advantageous to sell, hold, or buy the stock. The Known Pumper and Inflection point features serve as a binary decision to show where the tweet came from a popular pumper on Twitter, as well as whether or not an inflection point in the stock price occurred and if it was going up or down when it drastically changed direction. To validate our labeled training set, price was plotted against tweets where each tweet is colored by sentiment.

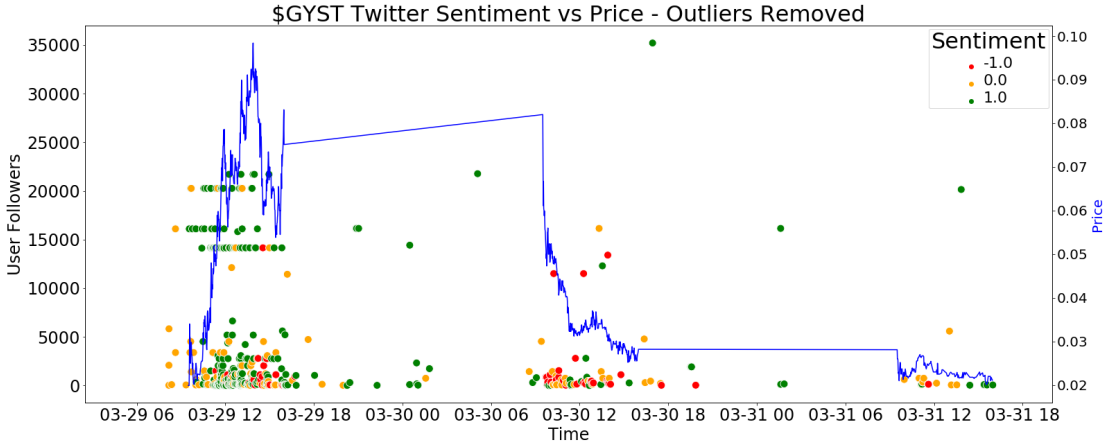

Figure 3: User Followers and Sentiment vs Price in known Pump

The sentiment shown in this case is reasonable considering a majority of the tweet volume and positive sentime occurs when the stock is first being pumped (denoted by green markings). However, as the stock price sharply decreases, the sentiment shifts to include more negative cases (denoted by red markings).

After generating the training set, the text of each tweet in the training set was passed through a word tokenizer to determine which words and emojis were most common for a particular sentiment scoring. An example of this step can be seen below in Figure 2.

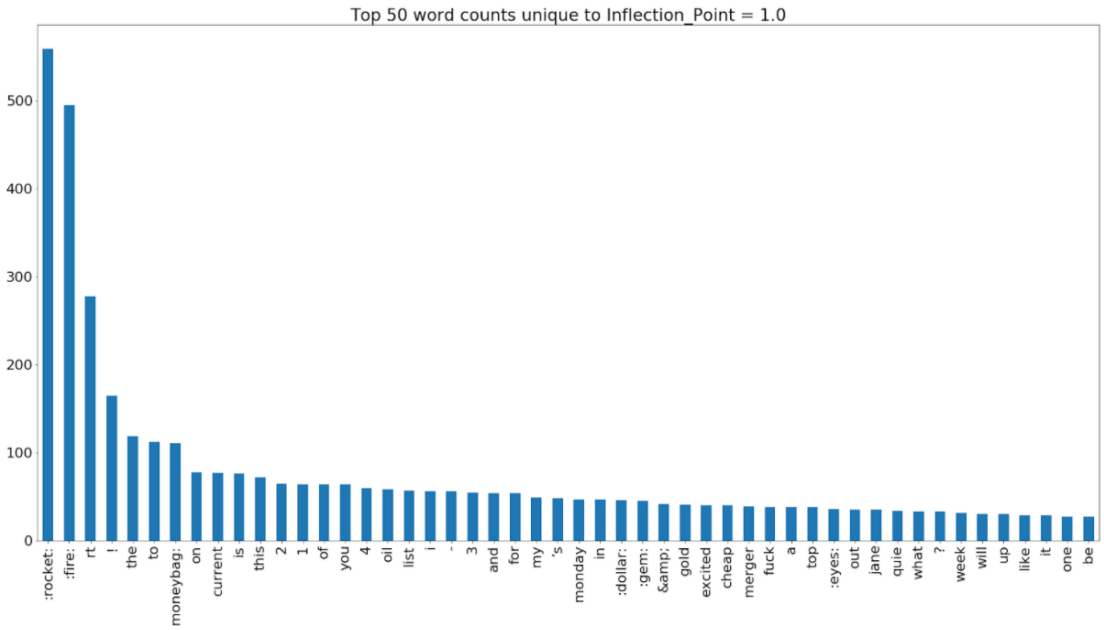

Figure 4: Top 50 Word Counts for Tweets near Positive Inflection Points

Figure 2 shows the results for the most popular words and emojis around a positive price inflection (aka the start of a pump). As seen, a high use of emojis such as the "rocket", "fire", and "moneybag" in tweets are a heavy indicator that the price is starting to head in a positive direction. Outside of emojis, words like "current" (as in an OTC stock upgrading to Pink Current status), "excited", and "cheap" all pass along tones of hope to potential buyers in the stock that do not want to miss out on what is viewed as an easy money maker. To counteract that, the most important words in tweets at the negative inflection point were observed as well.

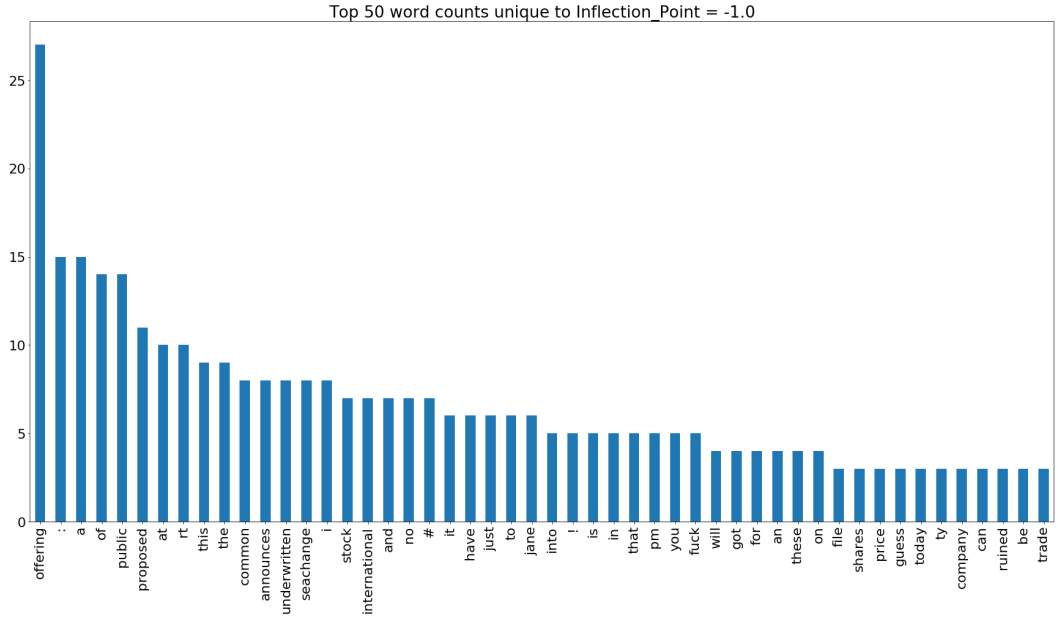

Figure 5: Top 50 Word Counts for Tweets near Negative Inflection Points

This exercise was done on the Inflection Points and the Sentiment Scores to provide popular words to use when determining the classification of tweets. Emojis were not as prevalent in the negative inflection points, but words such as “offering”, “proposed”, and “announces” stood out as common indicators of an impending negative shift in the stock price. These words seemed to correlate well with the old adage “buy the hype, sell the news”, with each word being related to either an announcement around the stock, or more specifically, a public offering of common stock.

Based on the analysis above, a list of pertinent words was created to generate additional features using a count vectorizer and TfiDf. The full list is available in ‘vocab.py’ but a sampling is provided below along with counts for a sample of tweets.

news	update	alert	company	2021	available	#stocks	free	:brain:	hold	:rocket:	:fire:	rocket	:moneybag:	buy	:pray:	:dollar:	moon	:bangbang:	:gem:
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0

Figure 6: CountVectorizer Example Words

Analysis

Next, the features generated from the count vectorizer and TfiDf vectors were used to fit various classification models with the intent of predicting sentiment on new tweets.

Sentiment Modeling

Multiple models were tested, including logistic regression, random forest, KNN, SVM, and Gaussian Naive Bayes. To select a model, the accuracy calculated from both a standard test set and cross validation set were compared across each model. An example of the test set accuracy comparison is shown in the figure below.

```
LogReg
0.8380758807588076
RF
0.8380758807588076
KNN
0.8177506775067751
SVM
0.8407859078590786
GNB
0.5047425474254743
```

Figure 7: Accuracy Results of Tweet Sentiment Predictions on Test Set using Various Models

Besides test accuracy, other factors such as precision and recall were considered in the model selection process. For example, many models had a high accuracy and precision but a low recall, particularly in the cases of negative sentiment. The figure below summarizes the results of a logistic regression model trained on the full labeled training set.

	precision	recall	f1-score	support
-1.0	0.91	0.34	0.49	238
0.0	0.88	0.10	0.18	236
1.0	0.87	1.00	0.93	2537
accuracy			0.87	3011
macro avg	0.89	0.48	0.53	3011
weighted avg	0.88	0.87	0.84	3011

Figure 8: Results of Logistic Regression Classification Model for full Training Set

In this case we see a higher accuracy and precision but a low recall for neutral and negative sentiment. This is primarily due to the fact that the training data is heavily skewed towards positive sentiment, a characteristic that is clearly evident in the confusion matrix below.

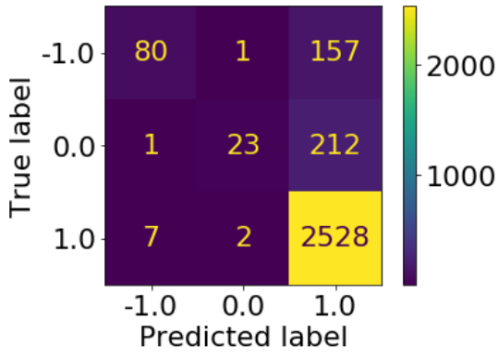

Figure 9: Confusion matrix from Logistic Regression Classification Model for full Training Set

To address this issue two approaches were investigated. First, a sample of the labeled training data was created where each sentiment classification was equally represented. Given that the training data only included 884 negative tweets out of 12,043 labeled tweets a sample of 884 tweets was taken for both neutral and positive tweets resulting in a new training dataset with 884 tweets of each classification (2,652 total tweets). When a logistic regression classification model was trained on this dataset, the overall cross validation accuracy dropped to 71% but the model performed better on neutral and negative sentiment classification. The resulting confusion matrix of the model trained on an even sampling of the data is shown below.

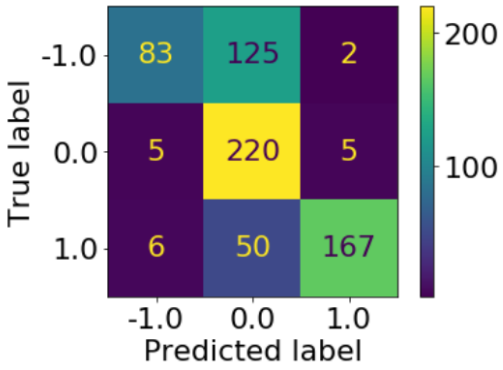

Figure 10: Confusion matrix from Logistic Regression Classification Model from even sampling

Per a suggestion from our adviser, another sampling approach was attempted where the negative and neutral cases were simply duplicated to match the number of positive tweets. Using this approach slightly improved the accuracy of each classification model and also created a more even distribution of predictions. For example, the random forest model improved from 70% to 72% cross validation accuracy. The confusion matrix for the random forest model is provided below.

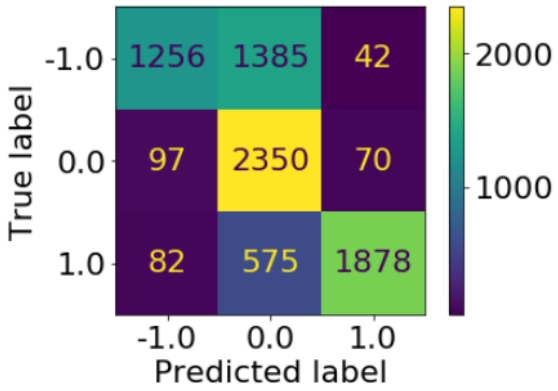

Figure 11: Confusion matrix from Logistic Regression Classification Model from even sampling

The confusion matrix above demonstrates an improvement in the predictions on neutral and positive tweets but still shows the model struggling to distinguish negative sentiment from neutral sentiment. Unfortunately, given the uneven distribution of the training data set, the sentiment model will most likely tend to predict more positive and neutral values until additional negative tweets are gathered.

Price Models

While the main focus of this analysis was to generate a sentiment model for tweets that reference stocks, a preliminary model was generated to predict price based on sentiment. First, the logistic regression model discussed in the previous section was applied to the remaining unlabeled tweets to generate a feature for sentiment in the data. Next, a simple linear regression was applied to features like previous price, number of followers, favourites count, and sentiment. Overall, this model was not very predictive in that it tended to lag behind the actual price by simply adjusting the prediction based on new data. The figure below provides an example of the model applied on the pump of GYST.

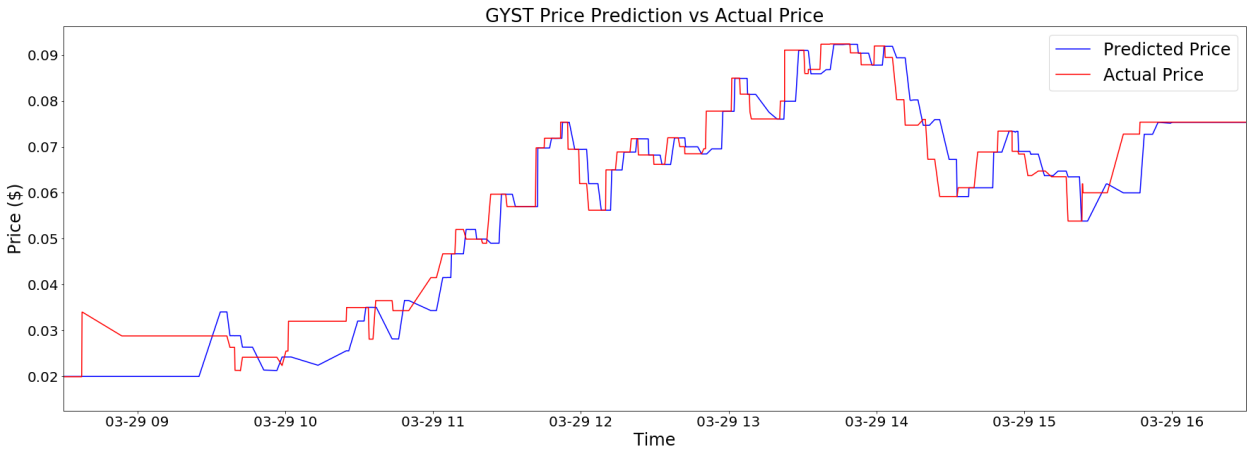

Figure 12: Preliminary Price Prediction Results

In this example the model was attempting to predict the price 30 minutes in the future given current price, user followers count, favourites count, number of emojis, and sentiment. Clearly, the model is lagging behind the actual price meaning it does not have enough information to anticipate future price

movements. However, when looking at the main features responsible for the models prediction, the top two features are current price and sentiment, indicating that sentiment is most likely the most important feature to pursue in future analysis.

Conclusion

Overall, this analysis demonstrates an approach to modeling Twitter activity with specific emphasis on sentiment. The sentiment models created from the manually labeled dataset showed some promising capability in classifying twitter sentiment on tweets that reference stocks. However, the main shortcoming of these models was a lack of negative tweets. Unfortunately, most tweets on stocks tend to be positive which is reasonable given the psychology of the situation. Simply put, it's more advantageous to be positive about a stock since this will likely lead to higher returns on your investment. Additionally, people on twitter may naturally be more positive and feel a greater barrier to posting a negative tweet.

Limitations and Future Exploration

Besides a lack of negative tweets, the restrictions from the Twitter API also made it difficult to gather tweets from historical examples of pump-and-dumps. The Twitter API enforces a strict 7-day limit on the number of tweets that users can pull, meaning analysis could only be performed on pumps that happened during the course of the project. With additional time, we could either gather more data as it becomes available or request a better license for the Twitter API that allows us to look farther back in time and generate more training data.

In addition to generating more training data, future exploration will likely include better predictive models on price. With the time limitation, this analysis only included a simple prediction on price using linear regression. Other better options may be neural networks, such as recurrent neural networks or even reinforcement learning strategies. Better models paired with the additional labels included in our training dataset such as tags for known pumpers and inflection points in price may allow us to generate better predictions on price.

In addition to future exploration plans, the lack of datasets in relation to stock-related tweets was another issue addressed during this effort. The dataset of manually scored tweet sentiment developed and used in these studies were uploaded to Kaggle https://www.kaggle.com/mattgilgo/stock-related-tweet-sentiment to be used by the broader data science community. The hope is that other data scientists will build off this dataset to identify pumps and give insight to newer and unaware investors prior to them falling for the various schemes through social media avenues such as Twitter.