Using Social Media Sentiment Analysis to Predict Stock Prices

Course project for CSE 6240: Web Search and Text Mining, Spring 2020

Marta Bras
Georgia Institute of Technology
martagaiabras@gatech.edu

Peter Butler
Georgia Institute of Technology
pfbutler@gatech.edu

Pedro H. R. Pinto
Georgia Institute of Technology
phpinto@gatech.edu

1 Abstract

The goal of this project is to explore the concept of emotional theory in the stock market by seeing if social media sentiment analysis can be used to predict a given company's stock price movement. To do so, we extracted, cleaned and generated sentiment for more than 3 million tweets on 54 companies and 4 million Reddit posts on 30 companies from the two main US Stock exchanges (NYSE and Nasdaq). To predict changes in stock prices, we fitted different machine learning models - Random Forest, Gradient Boosting, linear and Logistic Regression. We also tried different segmentation strategies; models per company, models per industry, models per cluster (after kmeans), and models for overall data. We tested different time lags - i.e. the number of days of data we use to predict a stock's movement. We conclude that there is strong association between social media sentiment and changes in stock prices. Random Forest had the highest precision on the Twitter dataset, while Gradient Boosting and regression perform well on the Reddit dataset. Additionally, fitting a model per cluster was the best strategy for the Twitter dataset, while an overall model was preferred for the Reddit dataset.

2 Introduction

Emotional theory in the stock market argues that stock traders don't always act in a rational and objective manner as the traditional stock market theories assert. Instead, stock traders are often driven by psychological factors and emotions. As a result, stock prices at a certain time might not be an accurate reflection of the underlying value of a stock but instead a collection of emotions from multiple players in a network.

The stock market being partially driven by emotions, indicates that there may be value in successfully extracting and give meaning to a network of emotions. The development of sentiment and network analysis techniques alongside machine learning predictive models and the widespread use of social networks, allows us to explore the concept of emotional investing in this project. We believe that this is a subject worthy of further research not only because it might be used by individual investors to maximize their trading payoffs but also because it might be used by companies to better forecast their future performance and day-to-day operations.

Understanding how emotions affect the stock market might have greater implications in terms of reducing overall volatility in the markets and consequently reacting better in moments of extreme panic. In that context, we also believe that the relevance of this project is not limited to the predictive models that we developed. The insights that we have gathered throughout the project, that we will briefly explore in this report, allow us to understand how emotions differ among companies, social media platforms and time periods and how the correlations between social media sentiment and stock prices are much higher than one might already expect.

We used a the PRAW and Pushshift APIs to collect daily Reddit posts on 30 companies for the past 12 years and GetOldTweets3 - a Python 3 library - to collect daily tweets on 54 companies for the last 3 years. We created two datasets and cleaned each tweet/ Reddit post by removing duplicates and missing data, decoding HTML

Conference'17, July 2017, Washington, DC, USA
2020 ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

text using BeautifulSoup, removing non-alphanumeric characters, stripping the text and converting all tokens to lower case. We then used the nlkt.sentiment package, a NLP library in Python that uses vader_lexicon to generate a sentiment score for each cleaned Reddit/Twitter text. The overall sentiment is scored in 3 categories - positive, neutral and negative - and a compound score (a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 and +1).

We combined each dataset with the daily changes in stock price for each company, which we collected using the IEX Cloud API. To combine the datasets, we aggregated the tweets/Reddits per company by various time lag periods and generated a mean overall compounded, positive, negative and neutral score. We also generated a weighted mean compounded, positive, negative and neutral compounded score - adjusted by the number of retweets/ Reddit score - for each company.

We built separate machine learning predictive models for Twitter and Reddit. Every model used the change in stock price with different time lags - from 1 to 15 days - as response variable and the compounded score for that day as the explanatory variable. For Twitter, we also included seasonality variables (week and month number) and post features (number of retweets and number of favorites).

We learned from our explanatory data analysis that the correlations between stock prices and social media sentiment varies greatly between industries and companies. Additionally, we envisioned that segmenting by industry might not reflect differences in correlations among companies and we decided to use k-means to create clusters of sentiment score - using mean compounded, neutral, negative and sentiment score per company as our features - to better reflect these differences.

We tested and compared results for models by industry (with all companies in the industry vs most relevant companies in terms of correlations), company, cluster and overall. We used precision as our baseline performance metric as we wanted our baseline to be a low risk investment strategy. We found that social-media sentiment associates strongly with stock returns. Random Forest had the highest precision on the Twitter dataset, while Logistic Regression performed the best on Reddit dataset. Fitting a model per cluster was the best strategy for the Twitter dataset, while an overall model was preferred for the Reddit dataset.

We conclude that the threshold selected for the trade-off between precision and recall reflects each investor's risk tolerance. The choice of threshold impacts the value of the predictive model and allows for different investment strategies.

3 Literature Survey

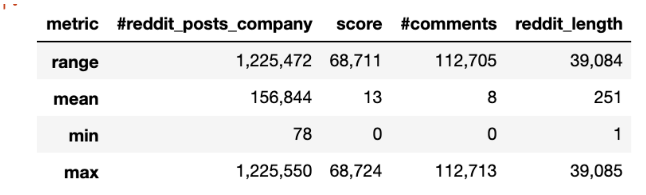
- **Sentiment Analysis of Twitter Data for Predicting Stock Market Movements**[2]: This publication explores a method for applying NLP and Machine Learning to a Twitter data set in order to predict movements in the stock market. An important difference is that a lot of this paper revolved around generating the sentiment scores for each tweet, while we used the Vader-Lexicon library and focused on testing different predictive models. The main shortcoming of this paper is that the authors used data from a single company (Microsoft), which can lead to over-fitting. In our analysis, we used a larger number of companies across different industries.
- **Correlating SP 500 Stocks with Twitter Data:** [3]. This publication delineates a more similar approach to our project by attempting to create a predictive model to a large group of companies based on Twitter sentiment analysis. An important difference is that this paper only collected data based on stock tickers (also known as "cashtags"), which reflect the sentiment about the companies' stocks and not how consumers feel about the companies themselves. Another important distinction is that we also incorporated data from Reddit and tested a variety of different models, while this paper only tested linear regression.
- **Social Media Sentiment Analysis For Firm's Revenue Prediction:** [1] In this publication, the author obtained data from both Twitter and Facebook, and applied a variety of models to predict a firm's revenue. Our linear regression baseline was based in one of these models. The main differences are that we are attempting to predict stock prices instead of revenues, and that we used the Lexicon-based approach as an input to our Machine Learning approach instead separating them.

4 Dataset Description and Analysis

4.1 Dataset description

Reddit data: We have collected data from 2008 to 2020 for 30 companies, with a total of 73 subReddits. The clean Reddit dataset has 4,705,320 rows and 17 columns.

The statistics for the most relevant features of the Reddit dataset - number of Reddit posts per company, score of each Reddit, number of comments of each Reddit and length of each Reddit, are summarized in the table below.



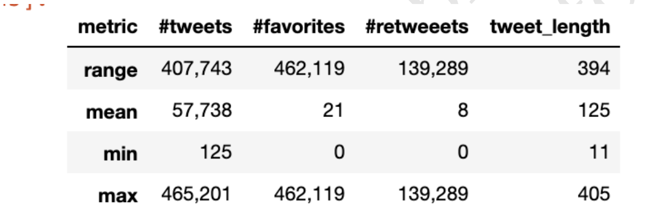
metric	#reddit_posts_company	score	#comments	reddit_length
range	1,225,472	68,711	112,705	39,084
mean	156,844	13	8	251
min	78	0	0	1
max	1,225,550	68,724	112,713	39,085

Figure 1: Statistics of overall Reddit dataset

Twitter data: We have collected data from 2017 to 2020 for 54 companies, with a total of 116 hashtags.

The clean Twitter dataset has 3,116,778 rows and 18 columns.

The statistics for the most relevant features of the Twitter dataset - number of tweets per company, number of times each tweet was voted as favorite, number of times a tweet was retweeted and length of each of the clean tweet, are summarized in the table below.



metric	#tweets	#favorites	#retweeets	tweet_length
range	407,743	462,119	139,289	394
mean	57,738	21	8	125
min	125	0	0	11
max	465,201	462,119	139,289	405

Figure 2: Statistics of overall Twitter dataset

4.2 Dataset analysis

The insights that we will explore in the following points were crucial to subsequently define our modeling strategies for the final part of our project. Our main insigths are:

1. There are differences in the level of social media sentiment generated among companies and platforms.

Some companies generate higher social media sentiment on both platforms, while others only generate higher social media sentiment on one of the platforms. Technology and telecommunication companies generate highest social media compound scores on both platforms, while commercial banks generate the lowest scores.

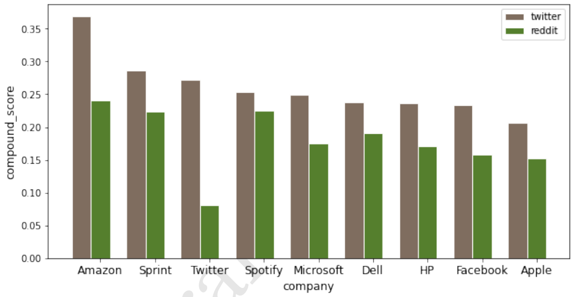


Figure 3: Companies that generate higher compound score on Twitter and Reddit

2. There is seasonality in the number of tweets/ subReddit pages and in the positive vs negative sentiment scores generated.

Over time, Twitter generates both higher positive and negative sentiment scores than Reddit. There are clear spikes in the sentiment generated, which seem to be similar in both platforms. The main spikes in the positive sentiment are generally around Christmas time. Additionally, the number of tweets and subReddits generated also vary greatly throughout the year. On average, the number of tweets is higher in the first 3 months of the year, dropping from almost 250k a month in March to 175k a week from April until the end of the year. The number of subReddits reaches maximum values from September to December and minimum values from April to August. We capture this seasonality in our models by using features that capture the time period - month and week number.

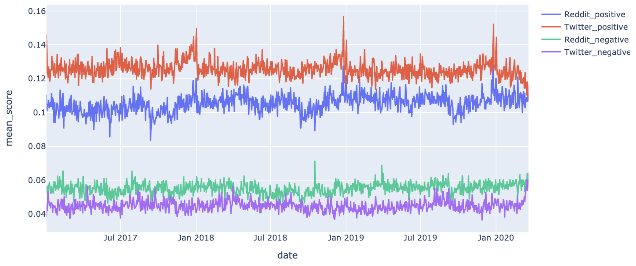
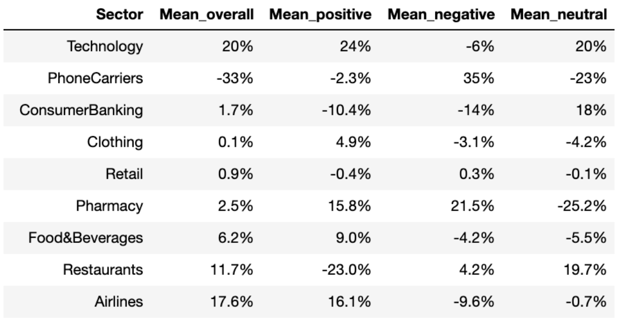


Figure 4: Seasonality in positive and negative scores for Twitter and Reddit

3. Some industries/ companies generate higher social media sentiment and have higher correlation with same day closing prices.

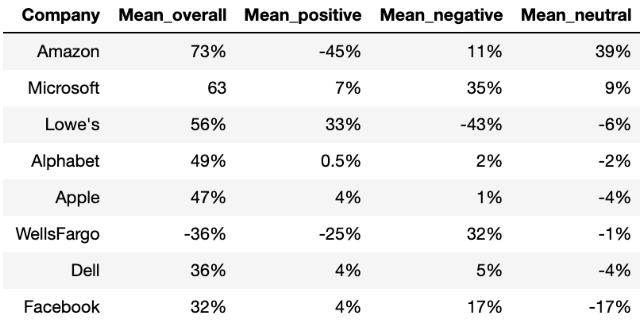
While Technology and phone carriers have high correlation with same day closing prices, all other industries have low correlations. We also concluded that compared to the weighted mean, the standard mean has higher correlations with closing price. Inconsistent directions in correlations might result from using same day closing prices when either the stock prices or the social media sentiment might take time to react.



Sector	Mean_overall	Mean_positive	Mean_negative	Mean_neutral
Technology	20%	24%	-6%	20%
PhoneCarriers	-33%	-2.3%	35%	-23%
ConsumerBanking	1.7%	-10.4%	-14%	18%
Clothing	0.1%	4.9%	-3.1%	-4.2%
Retail	0.9%	-0.4%	0.3%	-0.1%
Pharmacy	2.5%	15.8%	21.5%	-25.2%
Food&Beverages	6.2%	9.0%	-4.2%	-5.5%
Restaurants	11.7%	-23.0%	4.2%	19.7%
Airlines	17.6%	16.1%	-9.6%	-0.7%

Figure 5: Correlation between sentiment scores with same day closing price per industry - Twitter

The same thing is also true per company. The top companies are presented in the table below. Aside from a small number of other companies, the majority of the companies in the dataset have correlations that are lower than 10%. Additionally, we conclude that the mean compounded score seems to be the most useful sentiment score.



Company	Mean_overall	Mean_positive	Mean_negative	Mean_neutral
Amazon	73%	-45%	11%	39%
Microsoft	63	7%	35%	9%
Lowe's	56%	33%	-43%	-6%
Alphabet	49%	0.5%	2%	-2%
Apple	47%	4%	1%	-4%
WellsFargo	-36%	-25%	32%	-1%
Dell	36%	4%	5%	-4%
Facebook	32%	4%	17%	-17%

Figure 6: Top companies in terms of correlations to same day closing price

5. Using K-means clustering on the different sentiment scores we are able to capture clusters of companies that have similar sentiment scores for all metrics Clustering allows us to capture multi- dimensional sentiment scores as well as differences across companies.

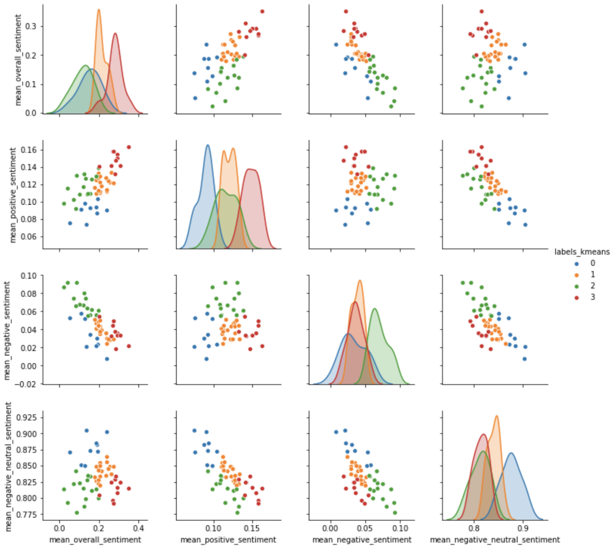


Figure 7: K-means clusters multi-dimensional visualization

5 Experiment Setting and Baselines

5.1 Experiment Setting

- **Data Split:** 80% for training and 20% for testing.
- **Cross-Validation Setting:** 3-fold Cross-Validation
- **Evaluation Metrics:** F1-score, Recall, Precision
- **System Settings:** 8GB RAM, 2 CPUs (2.39 GHz).

5.2 Experiment Baselines

The two baselines we used were Linear Regression and penalized Logistic Regression. Dimadi (2018)[1] showed that there was a correlation between Nike's social media sentiment and their quarterly revenue. Thus, we believe that Linear Regression may be a good baseline model to try for predicting stock prices of companies. Linear Regression works by fitting a linear model to given predictor to minimize the prediction error relative to a response. Since we are dealing with a classification problem (will a stock increase or decrease), we convert the response variable and our predictions to a binary indicator of positivity. We also used Logistic Regression as a baseline model. Logistic Regression is a common classification model that minimizes a logistic loss function instead of mean squared error, and thus may even be better suited than the Linear Regression procedure for such classification problems. We hypertuned the penalty term for this model using three-fold cross-validation. The penalty term shrinks the coefficients to prevent overfitting.

6 Proposed Method

We built separate models for the Twitter and Reddit datasets. The features in common to both models are enumerated below.

- **Response variable:** Changes in closing price
- **Models tested:** Logistic Regression, Random Forest, Gradient Boosting
- **Number of estimators:** 100
- **Hyperparameters:** Logistic Regression hyperparameter tuned using 3-fold cross validation. Random Forest and Gradient Boosting showed no signs of overfitting, so they were trained with no minimum impurity split or max depth constraints.
- **Segmentation strategies:** Models by industry and company

Twitter model exclusively: We performed a clustering analysis to better capture differences in sentiment scores across companies and sentiment metrics. Our clusters were created using k-means clustering with the the mean compounded, positive, negative and neutral scores for each company in the Twitter dataset as our features. Additionally differences include:

- **Explanatory variables:** We used post features - mean compounded daily score per company; - user/post features - number of favorites, number of retweets; and **seasonality features** - week number and day number
- **Segmentation strategies:** Segmentation by clusters

Reddit model exclusively: We engineered additional features including a weighted sentiment score which is weighted by the log of the respective Reddit post score. For predicting stocks, we look at specific time frames of Reddit activity and take an averaged weighted score (for compound, positive, negative, and neutral sentiments) as well as proportions of positive and negative posts and the total number of posts over

Improvements from baselines:

- More flexible machine learning classification models - Random Forest and Gradient Boosting vs regression.
- Segmentation strategies: segmenting by company, cluster, and by industry.
- Cross validation to select optimal lookup period.

7 Experiments

7.1 Baseline results

The baseline results are summarized in the following table. We used the same experiment setting to run our proposed models.

Table 1: Baseline Results

Dataset	Model	F1 score	optimal lookup
Reddit	Linear	0.687	10 days
Reddit	Logistic	0.686	15 days
Twitter	Linear	0.668	15 days
Twitter	Logistic	0.683	6 days

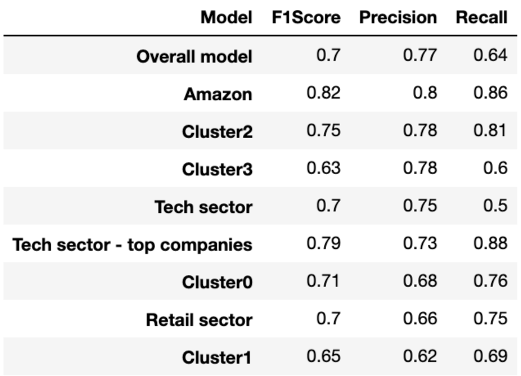
7.2 Proposed method - results

Twitter For the Twitter dataset, Random Forest with 14 day lag in changes in closing price had the highest precision for almost every segmentation strategy. The model using only data for Amazon was by far the one that performed the best, with a precision of 0.8 and a recall of 0.86.

The results for all strategies are displayed in the table below. We ordered our results in terms of precision, as we want to have a low risk investment strategy as our baseline. By precision, cluster 2 and cluster 3 are the ones that perform the best after Amazon. Cluster 2 performs well in terms of recall and overall F1 score as well, which means that it can be used not only for defensive but also offensive investment strategies. The tech sector only including top companies has a great recall - good at identifying buying opportunities - but it also captures a lot of false positives (signals stocks as going up when in fact they go down).

Reddit For the Reddit data, we got the best results by training the models was on all data (without segmentation). Linear and Logistic Regression had great precision but low recall. These models performed eqaully well no matter the number of lookback days and their results were about the same for each individual company. The Random Forest performed worse overall than either of the baselines. Gradient Boosting's performance does not best the baselines overall, but it does much better for predicting the performance of certain companies.

For example, the boosting model predicts Fox Corporation's performance with 0.93 precision and 0.73 recall.



Model	F1Score	Precision	Recall
Overall model	0.7	0.77	0.64
Amazon	0.82	0.8	0.86
Cluster2	0.75	0.78	0.81
Cluster3	0.63	0.78	0.6
Tech sector	0.7	0.75	0.5
Tech sector - top companies	0.79	0.73	0.88
Cluster0	0.71	0.68	0.76
Retail sector	0.7	0.66	0.75
Cluster1	0.65	0.62	0.69

Figure 8: Performance metrics for the different segmentation strategies fitting Random Forest with 14 days lag

Table 2: Reddit Results

Model	F1	Precision	Recall
Linear	0.686	0.961	0.530
Logistic	0.687	1.0	0.522
Forest	0.538	0.553	524
Boosting	0.628	0.734	0.549

7.3 Comparison Proposed vs Baseline

Twitter:

For Twitter, our overall model performs better than baseline - 0.68 vs 0.71 F1-score. We compare the results from the baseline vs our model in the following ROC curve.

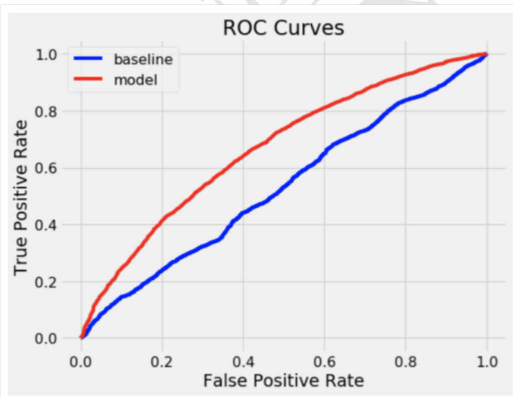


Figure 9: ROC curve - baseline vs our proposed model

Reddit The baselines outperformed our proposed models overall. However, the Gradient Boosting does achieve much better recall at the cost of some accuracy for many of the individual companies. The Gradient Boosting classifier is certainly more flexible, but for the purposes of developing a conservative trading strategy, Logistic Regression seems to do a much better job.

8 Conclusions

The main limitations in our models are related to the data extraction and data cleaning process - extracting the necessary data from Twitter is a lengthy process that in our case lasted for more than 4 weeks just to get 3 years of data. In the data cleaning process, more focus should have been given in excluding promotional and company-owned tweets. In terms of modeling, a key limitation was not using neither company fundamentals nor market features, such as market risk premium, company size, market-to-book value and total assets.

Future work should focus on getting more historical data for Twitter and being more efficient in excluding irrelevant tweets. In terms of modeling, future research should test models with additional features and even test high frequency trading by developing models with real time data. Our work can be used as a starting point for the development of investment strategies. The threshold between recall and precision impacts the choice of predictive model and reflects investors' risk tolerance. In the future, short-selling vs long strategies can be developed based on investor's risk tolerance and consequently choice of predictive model.

9 Contributions

All team members have contributed equally.

References

[1] I. Dimadi. 2018. SOCIAL MEDIA SENTIMENT ANALYSIS FOR FIRM'S REVENUE PREDICTION. In *UPPSALA UNIVERSITET. Upsala, Sweeden.*

[2] Sasank Pagolu, Kamal Challa, Ganapati Panda, and Babita Majhi. 2016. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. *International conference on Signal Processing, Communication, Power and Embedded System* (2016).

[3] Bing Wang Yuexin Mao, Wei Wei and Benyuan Liu. 2012. Correlating SP 500 stocks with twitter data. *Workshop on Hot Topics on Interdisciplinary Social Networks Research.*