Predicting the Stock Market with Twitter

Corey Clemente
coreycle@bu.edu
Boston University, Department of Computer Science
111 Cummington Mall
Boston, MA 02215

Renzo Callejas
rcallejas@bu.edu
Boston University, Department of Computer Science
111 Cummington Mall
Boston, MA 02215

Abstract

The main idea of this project is to answer the following question: can one accurately predict stock market movements using Twitter data extracted and analyzed using data science techniques? We filtered tweets using stock symbols and extracted and analyzed the tweet sentiment, date, and popularity. We also mined stock data from Yahoo Finance and aggregated the stock price and the date. Using these two datasets, we used data science techniques to try to find the existence of a correlation between the stock market and sentiment on Twitter. We expect to find a very small correlation because of the randomness of the stock market as well as the other important factors that affect the stock market not related to social media.

Introduction

The main motivation for our study is to find a more accurate way to predict the stock market. This study is relevant for financial theory academics searching for previously unknown variables that cause market shifts. Applications also include people in the financial sector looking to increase the returns of their investments using a publicly and easily accessible, live dataset.

We tried to answer two questions:

1. Can stock movements of specific companies be predicted with Twitter?
2. Can overall stock market movements be predicted?

We created a dataset of 2015 tweets from Twitter and another dataset of stock prices over the same time pulled from Yahoo Finance.

Our main approach was to pick eight companies, and get the Twitter and stock data for each company for the year of 2015. The Twitter data for company A would be tweets that mention that company. The features extracted from those tweets would be sentiment and popularity. The stock data for company A would describe if the company's stock went up or down for each day of the year. Then, using classification such as Decision Trees, Support Vector Machines and Linear Discriminant Analysis, we investigated the existence of a correlation between tweet sentiment and stock market movements. In other words, do the Tweets about company A on Monday predict how that company's stock will do on Tuesday? We then took all the tweets about all the companies and tried to predict the stock movements of each individual company. In other words, do all the tweets about companies A, B, and C predict the stock movements of each individual company?

We expect for the experimental results to show a very small correlation between tweets and the stock market.

Techniques

We will use the following classification algorithms: K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forests, Adaboosting, Gaussian Naive Bayes, Linear Discriminant Analysis, and Quadratic Discriminant Analysis.

These are the standard classification techniques used to learn about vectorized, integer-valued data and attempt to classify it into two buckets: a stock increase or a stock decrease. Since we are predicting it will be difficult to find a correlation due to the stock market's randomness, here we attempt to utilize different style classification algorithms to see what works best. Statistical, decision and space-based classifiers are all very different and may provide certain insights into the data collectively that we would otherwise not attain by using fewer.

We predicted that SVM using a radial basis function would perform best, since it transforms the data into a higher (sometimes infinite) dimension that can be used to find some correlation that can be separated by a hyperplane. This type of classification seemed strong considering the low amount of correlation we predicted we could find. K-Nearest Neighbors might also perform well, trying to find similar occurrences of tweet sentiment and stock movement in the past (our dataset) and applying that logic to future predictions, however since the stock market is predicted to be so random, trying to find similarity in the past (using a space-based method) might be too simple of an attempt to learn from the data.

Random Forests also appeared to be a strong method to test, since it is comprised of multiple Decision Trees given randomized (then learned) weights. Again, given the type of data, the complexity might be able to find deep structure.

Datasets and Experiments

We had one dataset with Twitter data and another dataset with stock data from Yahoo! Finance. We scraped Twitter.com to get 400 tweets per day for the year 2015, for each of eight chosen companies. We filtered the tweets by using the company name as the keyword. In total we had a dataset of 1.2 million tweets. We did not use the Twitter API to get our data because the API doesn't allow for historical lookups. We chose to analyze the following companies: Citigroup, Netflix, Tesla, Twitter, McDonalds, Walmart, Microsoft, Disney. Each company had a corresponding pickle file that stored data from tweets mentioning that company. The data included the following: the tweet, the number of retweets, the number of favorites, the user, and the date. Using two sentiment analysis algorithms, we then got a sentiment rating for each tweet. Using NLTK we got a compound, positive, neutral, and positive rating. Using TextBlob we also got polarity and subjectivity. We added all of these sentiment scores to our data.

The following describes the stock data. For each of the eight companies, and for each day that the stock market was open in 2015, the following data was listed: date, 1 if the stock went up or -1 if the stock went down, and the corresponding number of points the stock went up or down. The change in the stock was calculated by taking the stock price at the end of the day and subtracting it from the stock price at the day's open. We used the Yahoo! Finance API to get this data.

In our initial round of experiments, we used the following techniques individually: K-Nearest Neighbor, SVM (Linear and RBF), Decision Tree, Random Forest, Ada Boosting, Gaussian Naive Bayes, LDA (Linear Discriminant Analysis), and QDA (Quadratic Discriminant Analysis). We split the data into a 70/30 for testing vs. training. We took the testing data randomly, created the necessary vectors, and put the same random data into each of the models so that we could compare their performance.

In our second round of experiments, we created a Random Forest as a classifier that would be trained on how each model performed in the initial round. The idea is to see if we could learn when certain models were correct vs. when certain models were incorrect so that we could increase our accuracy.

In our final round of experiments, we used our own heuristic by placing weights on each feature, and repeating the experiments of round one and two. We hypothesized that tweets with a higher popularity might be more influential in changing the stock market. Based on that hypothesis, we experimented with giving tweets that had higher retweets and favorite value a higher importance value in our models. We also hypothesized that tweets that were very subjective were better predictors of consumer opinion and sentiment and therefore the stock market (as opposed to tweets that were from industry professionals that usually had low subjectivity scores). To test this we gave tweets that were more subjective higher weights. For both of these hypotheses', we reran our first two experiments. The results were only better by a fraction of a percent, and therefore they were still very close to random and did not have a large predictive power.

Results and Discussion

Upon training each model for each company, and models for all companies combined, we attained the following results:

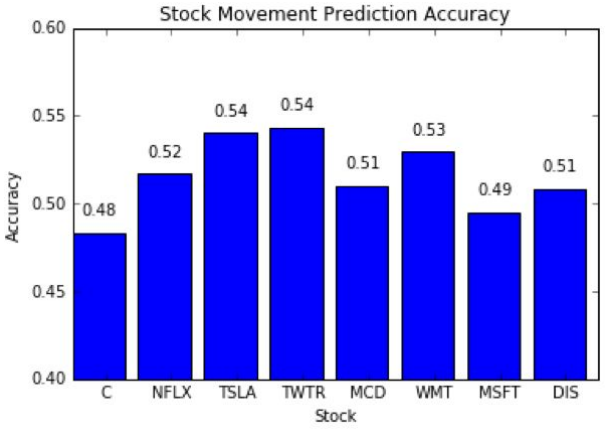

Figure 1 - best prediction rate per company

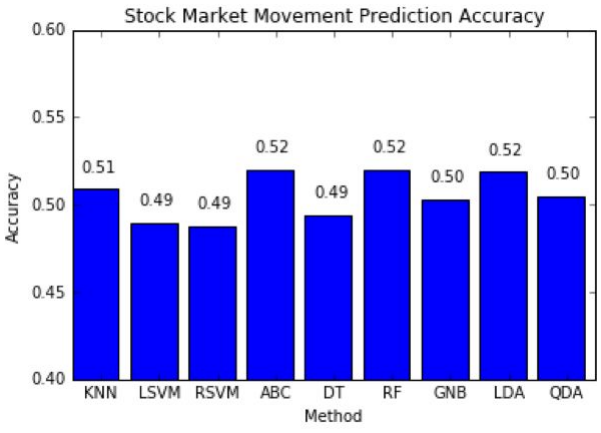

Figure 2 - model performance

In Figure 1, we see the results of the best prediction attained for each company. Twitter and Tesla tweets have the highest predictive power at 54%, meaning that little correlation could be found between Twitter sentiment/activity and stock movement. Each of the eight chosen companies fell between 48-54%.

Essentially, these graphs show that, using our data, there is little difference between randomly guessing whether a stock would go up or down and using trained models.

Figure 2 shows a similar results. Combining each of the companies into a “mini” stock market, and using all the gathered tweets to predict a general increase or decrease in stocks, each model also falls in the 49-52% accuracy range. Adaboosting, Random Forests and Linear Discriminant Analysis performed best with the data, but again, results would be similar to randomly guessing.

Overall, our results agree with our hypothesis. The randomness of the stock market is easily visible with these numbers. For example, using Walmart tweets to predict Walmart’s stock performed nearly identically to using all tweets to predict their stocks. Even preprocessing the data in different ways, such as assigning an importance value to a tweet and adding more layers of sentiment analysis, provided little difference, randomly increasing or decreasing the accuracy of our models by 1% at most.

Conclusion

In conclusion, we found very little to no correlation between twitter sentiment and stock movement using the data we were able to collect. Using our models, an investor would only perform slightly better than if she were to invest randomly. We believe the biggest reason why our models don’t have a higher predictive power is because the Twitter data needed to be cleaned extensively in a way that is backed by sound economic thinking. For example, if one thinks that consumer behavior is more predictive than company behavior, than the Twitter data should be cleaned to get rid of all tweets that aren’t by individual Twitter users who are expressing opinions about a product. Company tweets, ads, and news tweets should not be included in that model. We think the biggest problem is that the Twitter data we used represented a field of opinion that was simply too large and not weighted by importance of the Twitter user. Cleaning data in this way will take much more time and diligence, and is something we will do when we continue this project during the summer.

We also plan to get a much larger Twitter dataset to work with, which will increase the confidence of our findings. Our dataset of 1.2 million tweets was good for this project, but not large enough to give a definitive statement on our results as they could have been based on tweets that were not representative of the entire set of tweets that Twitter holds.

Additionally, we plan to create a sentiment analysis algorithm that is unique to Twitter and deals better with the type of grammar and vocabulary used by users on Twitter, and that can also deal with emojis. We used the most well known sentiment analysis algorithms, but they do not perform very well with Twitter data. Based off of looking at Tweets and their respective sentiment scores, we could tell the algorithms simply did not have the capacity to reflect the true sentiment of the Tweet. By creating our own list of positive and negative words tailored to Twitter, we predict the model will have a higher predictive power.

In conclusion, our finding of a small correlation between Twitter sentiment and the stock market are what we expected to find. This project reinforces the efficient-market hypothesis that claims the stock market is a “random walk” and will be hard or impossible to predict with a reasonable level of accuracy.