Predicting the Stock Market with Twitter

Corey Clemente, Renzo Callejas · Professor Kollios

Introduction

The stock market is difficult to predict, and therefore it is challenging for investors to make informed investment decisions. Can social media be used to help solve this challenge? Specifically, can public tweets from Twitter reflect average investor sentiment, and therefore predict changes in the stock market?

Goals

The main goal of this project was to find if, using prediction techniques learned in class, changes in the stock market could be predicted using tweets.

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

Data

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.

Preprocessing

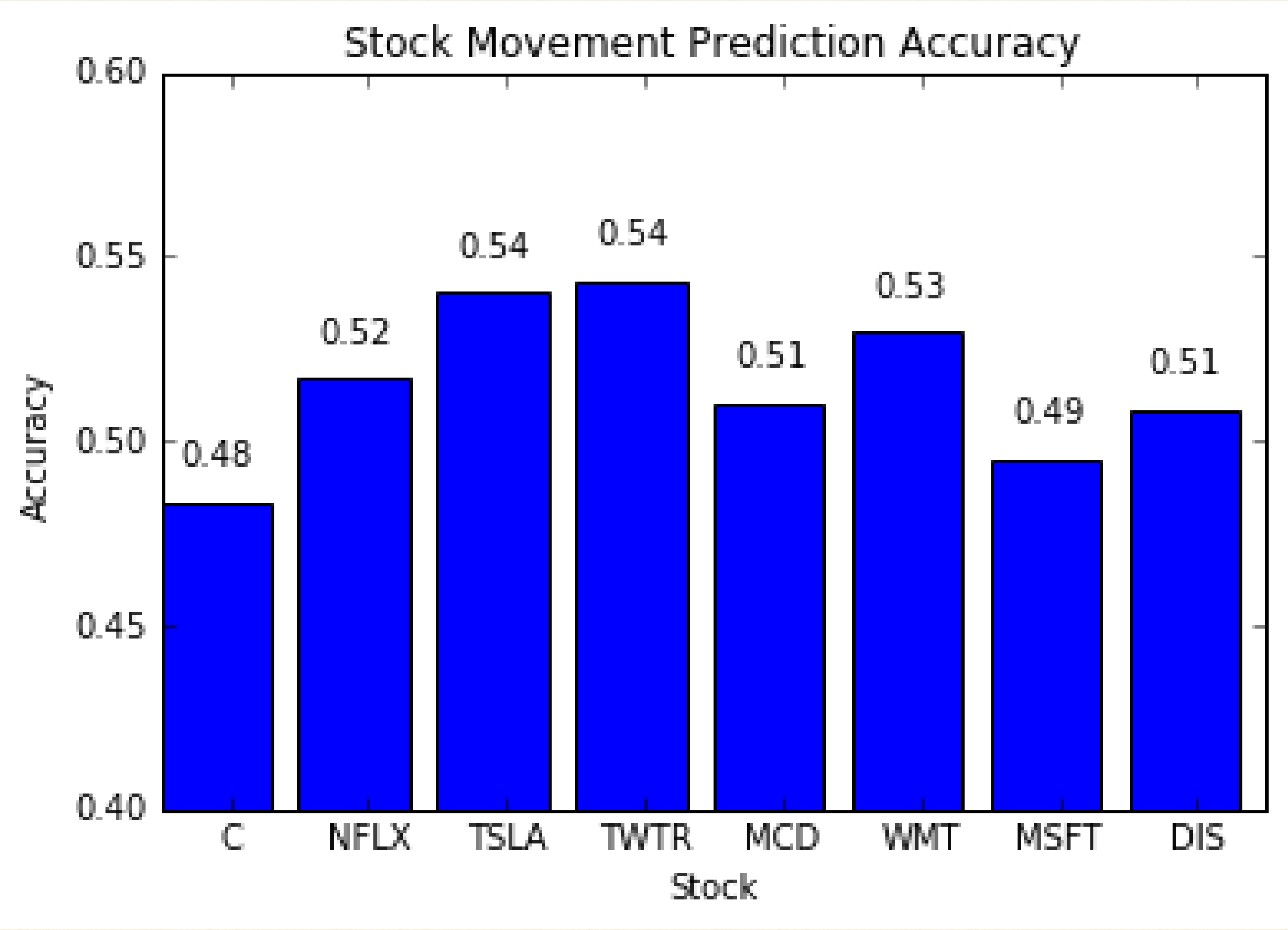
We created a vector for each company that had average values for each day of the year. The elements were the sentiment of the tweet, the number of retweets, the number of favorites, and the previous day's stock change.

Companies

We chose to analyze the following companies: Citigroup, Netflix, Tesla, Twitter, McDonalds, Walmart, Microsoft, Disney

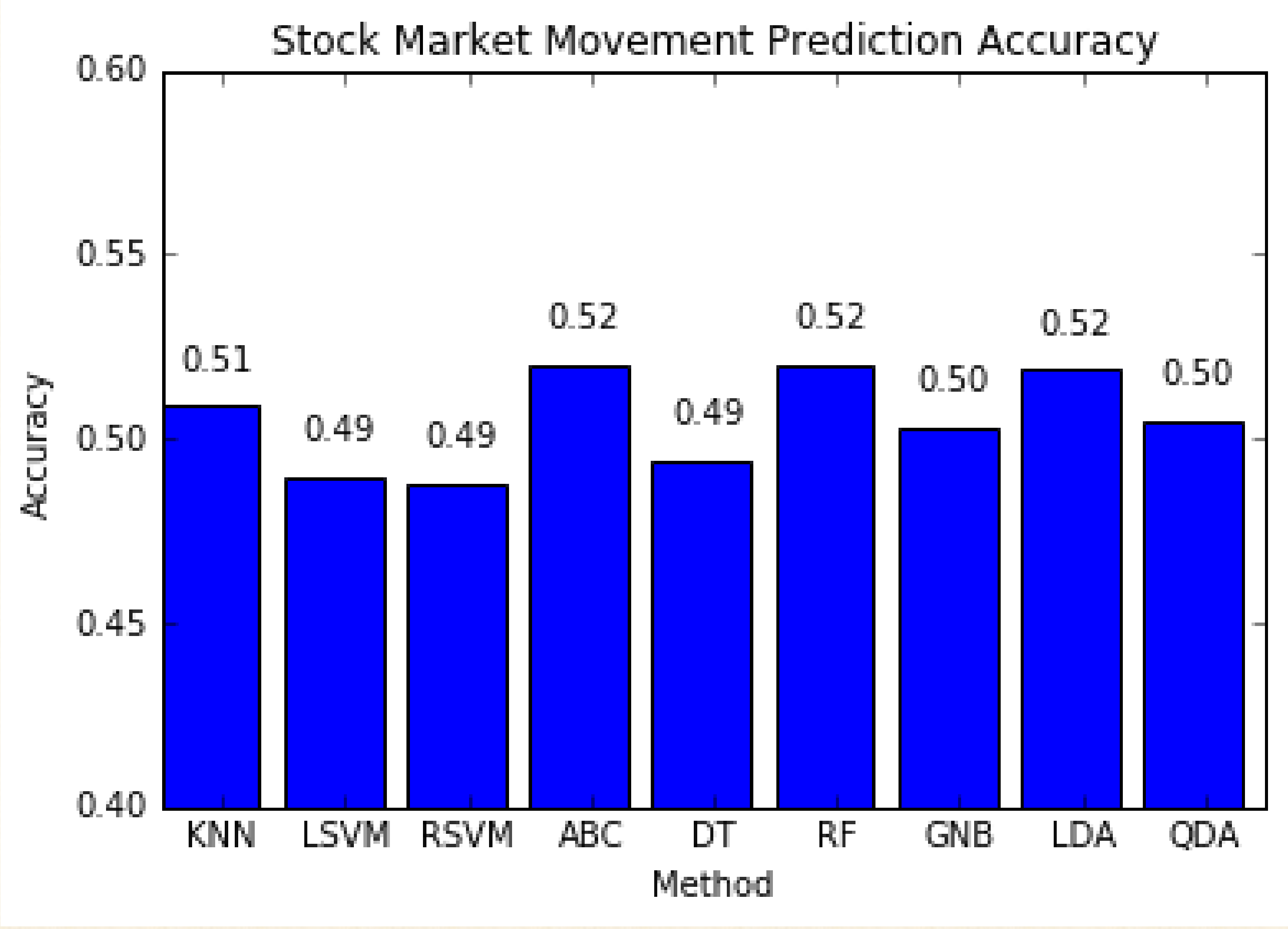
Predictions for Eight Companies

The following graph shows the accuracy of our model on eight different company stocks.



Prediction for Overall Stock Market

We predicted the overall movement of a "mini" stock market that included all eight companies, and plotted the results based on prediction algorithm.



Methodology

Attempted to predict increase or decrease in daily stocks using previous day's tweet data

70/30 random split for testing vs. training

Used following models:

- K-Nearest Neighbor
- SVM (Linear and RBF)
- Decision Tree
- Random Forest
- Ada Boosting
- Gaussian Naive Bayes
- LDA, QDA

Created our own heuristic by placing weights on each model to try and improve accuracy

Results

Across each company, average prediction accuracy was **51.5%.**

Twitter and Walmart stocks were most predictable at **54%** and **53%**, while Citigroup was least predictable at **48%.**

Random Forest classifiers and SVM (using a radial basis function) performed best.

Conclusion

In conclusion, we found very little to no correlation between twitter sentiment (after keyword extraction) and stock movement using the data we were able to collect. While partly expected, this project tends to agree with standard economic thinking that the stock market is a "random walk" and will be hard or impossible to predict with a reasonable level of accuracy.

Future Work

Get more data.

Be more selective with data, i.e. select public information from financial professionals.

Better feature extraction from text - find out what (if any) information could be better predictors other than sentiment and previous history.