SSC394E Final Project: Relations between Twitter and Stocks

By
Suvamsh Shivaprasad

Introduction

Communications methods have widely improved over the years. Today a large portion of communication happens over the Internet, specifically through social media. Social media giants such as Facebook and Twitter have been able to collect extremely large volumes of data. Twitter users send over 400 million tweets per day. Many of these tweets would have a great relation to consumers' sentiment towards different companies. Sentiment analysis of users' tweets could be used to predict the fluctuations in the stock market.

Twitter has been known to provide very quick and reliable information. It is very well implemented in high frequency trading (HFT). HFT firms need information that arrives very quickly and accurately. Unlike twitter, newspaper articles may take hours if not days to be written, which is a large disadvantage in a world of nanoseconds.

Method

For this project I have chosen to work with three large corporations with consumer facing products: Microsoft (MSFT), Apple (AAPL), Google (GOOG). I specifically picked these companies because they have disruptive products that cause social media uproar. I was able to query Twitter's public feed with words related to the companies. I pulled daily tweets related to the company and applied sentiment analysis. From there I was able to discover whether a tweet had a negative, neutral or positive sentiment. Concluding that this sentiment is towards the company I was able to build a model to see the growth of sentiment over time. The financial data was pulled from Yahoo using the Yahoo Finance API. This allowed me to get an open price, close price and adjusted close prices from the date of IPO.

Preprocessing

Filtering

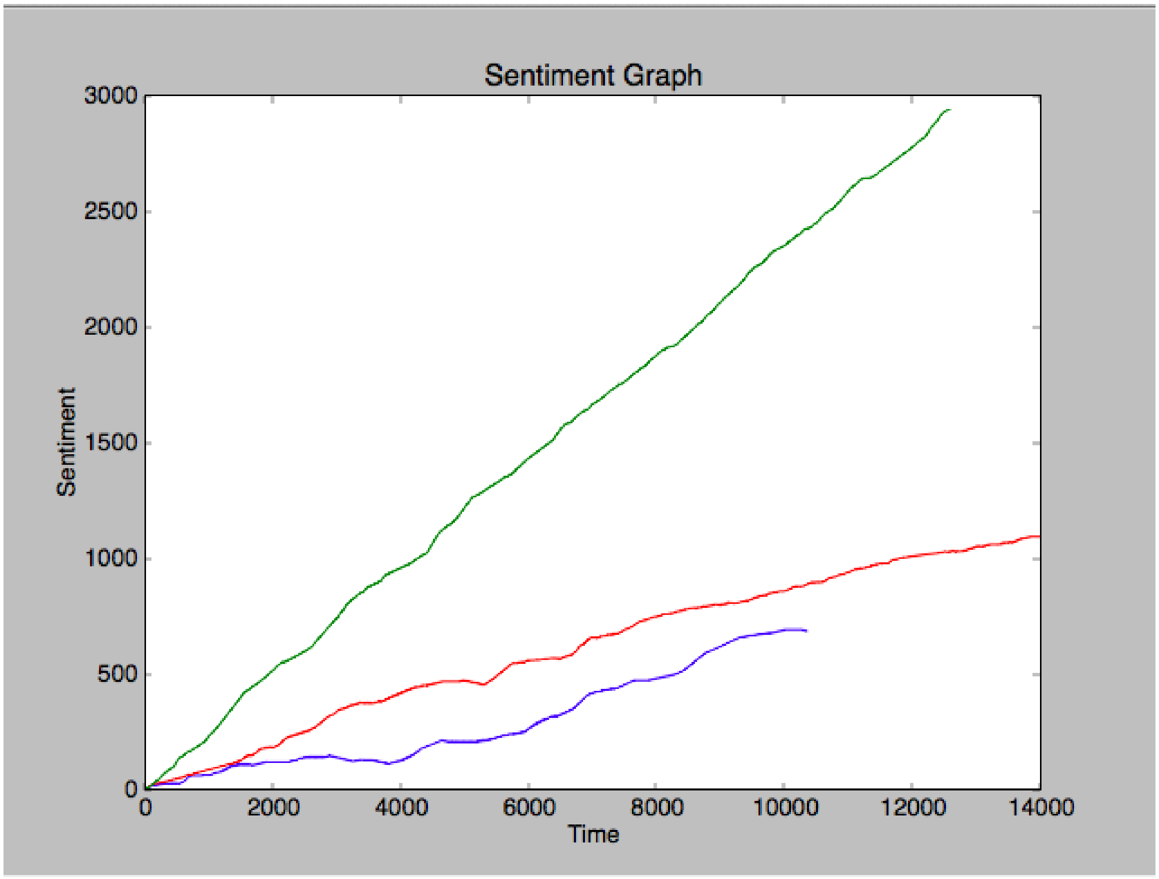
There were a lot of redundant tweets not related to any of the three companies. I first needed to filter them out and only work with data related to the three companies. By selecting specific key words that are closely related to the company, for example words related to Microsoft include windows, azure, msn, Xbox, etc., I was able to filter the public data set.

Clean Data

Cleaning the data is essentially two-step process. First I needed to classify whether a tweet is noise or useful information. Using a named entity recognition system to identity if the tweets have entities related to the company can do this. Then I needed to normalize the tweets. Tweets contain phrases like "soooo cooool duuude" which can be normalized to "so cool dude". This is the second method to clean the data.

Sentiment Analysis

The sentiment analysis was done using Sentiment104.com free API. I uploaded a JSON file containing all the tweets along with specific query words. A JSON file was returned adding a polarity parameter to each tweet. The polarity returned is negative, neutral, or positive. Using this I could gauge the sentiment of the public. Below is a graph of the sentiment towards Google (Green), Apple (Red) and Microsoft (Blue):



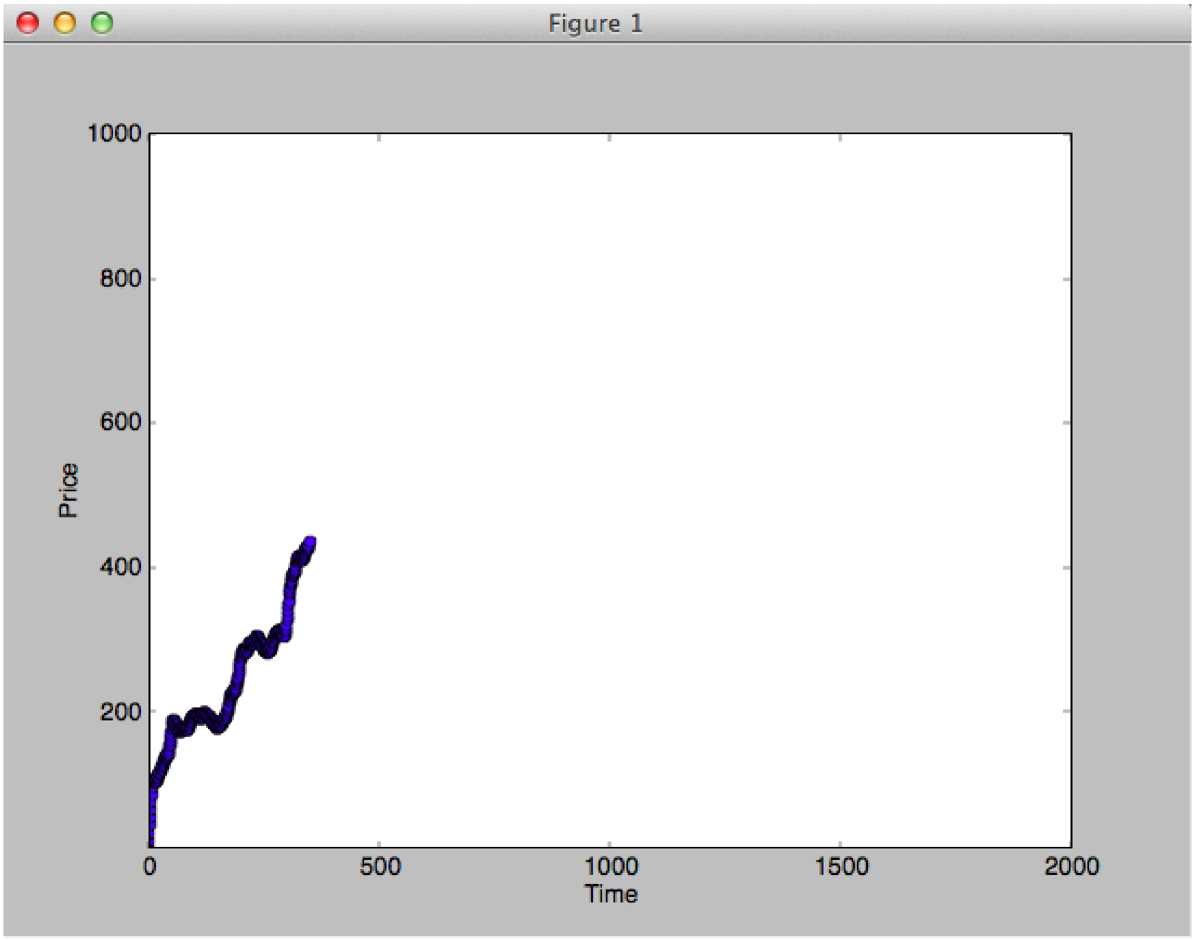
We can see here that all three companies seem to have good relations with their customers. Take for example Microsoft, who just recently released their much awaited Xbox One console. By observing these tweets I can say a good number praised the Xbox One. This is why we can see the growth of sentiment towards Microsoft grow positively. Similar situations cause Google and Apple's sentiment to have positive growth.

Prediction

Using the above sentiment analysis data I could make predictions. Again taking Microsoft for example, as there was a growth in positive sentiment I could conclude that it would be a good time to buy the stock. If live tweets were used it would be able to beat the change in market.

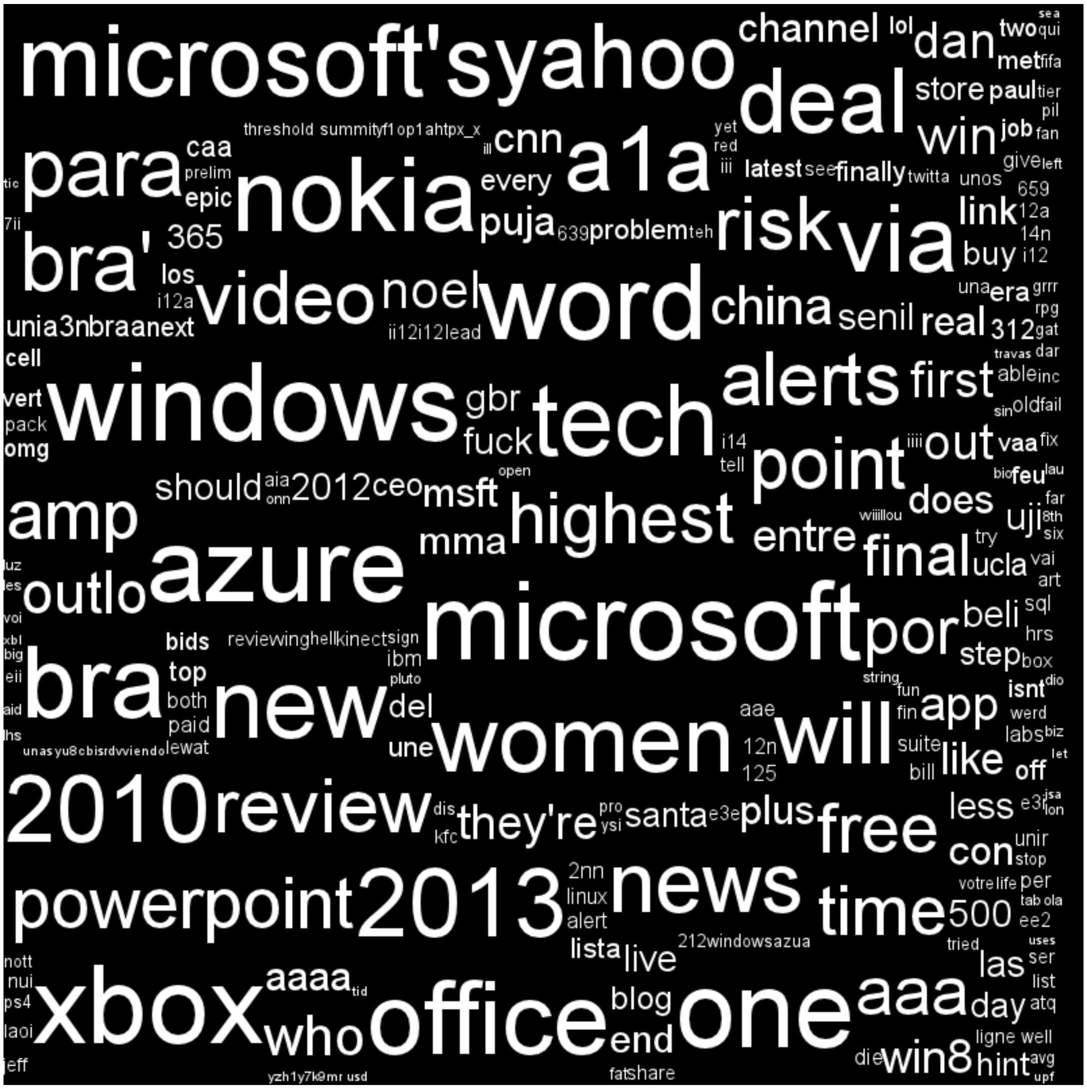
Historical data

I was able to plot the growth of a company's stock from IPO on an animated graph. Then I calculated a moving average across the graph to smoothen it out. A screenshot of this feature is provided below:



Word Cloud

A great way to visualize large amounts of textual data is through word clouds, in which words are scaled to size with respect to their frequency.



Conclusion

Twitter is a great source of immediate news, but using it as the only factor in a trading model is not recommended. Sentiment analysis cannot be the only argument while constructing a trading model. In industry today, Twitter is only one dimension in most multi dimensional complex trading models.