Sentiment Analysis of Twitter – Midterm Project Report

Christine Mayer clm293

Project Goals

Over the course of the semester, I aim to develop a model to predict future stock market data based on a sentiment analysis of Twitter data. The hypothesis is that data encapsulating how a certain publicly traded company is being discussed – that is, whether people online tend to be talking positively or negatively about the company and its products – will provide insight into the performance of the stock in the immediate future. The major steps for my project are as follows:

1. Gather data and analyze sentiment
2. Implement a time series model to predict the next time point
3. Verify the model
4. Compare results with a machine learning method

Current Progress

Thus far, I have set up a system to pull live data from both the stock markets and from Twitter to construct a set of recent data for my model.

In selecting stocks to analyze, I have chosen companies with products and services provided directly to individual consumers. I predict that these types of companies will be mentioned more frequently in social media in a context that would indicate customer satisfaction, and thus correlate to performance. Although I selected 5 stocks to consider for right now -- Microsoft, Google, Verizon, Burger King, and McDonald’s, this set is likely to evolve as I attempt to optimize my results and look for interesting patterns.

As I collect Tweets using the Python package Tweepy, I perform a series of Natural Language Processing related tasks before ultimately deriving a sentiment score for each mention. I generate lists of the most recent Tweets with an appropriate key word for the stock, and then run some preprocessing on the texts including removing URLs and deleting duplicates (Retweets) that might skew data. For now, I am using the Textblob library to assign a sentiment score to each Tweet, and an average score for the current set of data. As I continue to work, I plan on experimenting with different sentiment analysis tools. Currently, I am looking at changes to the histograms of the Twitter sentiment data while watching for correlated changes in stock prices.

The Twitter API restricts the data available to pull and mine to the fairly recent past, so I am in the process of developing a time series data set that collects the day’s relevant tweets and analyzes the sentiment. This is what I will compare to daily stock percentage changes.

Future Plans

My next step is to fit my gathered datasets to a time series model. Here is where I aim to predict the next time-point in the stock market data by interpolating how Twitter sentiment has affected stock percentage changes in the recent past. I plan to experiment with several models, including the basic ARMA model, and then moving on to try ARIMA and GARCH models. I will use least squares regression to fit models to the data to the models, then evaluate my results.

The next step will be to verify my models to see how effective the results are. I plan to further research some of the popular verification tools, including Box-Liung and McLeod Li, as this is a crucial step in proving the legitimacy of the model.

Finally, I will compare my models’ performance to other strategies leading the competition in this area – mostly ML methodologies. I will look at the performance of neural nets and deep learning in this task.