Get Rich - Generalized Examination of Tweets for Recommendations of Investment and Stock Purchasing

Jonathan Sumrall, Sander Kools, Giedo Mak, and Michiel Fortuin

Technical University of Eindhoven

Goal

Using a general sample of tweets, calculate daily mood trends and predict future prices of key stock market indexes.

Data

We collected tweets from Twitter, who provide a stream of tweets as they arrive. Historical tweets from Archive.org were used for testing and development.

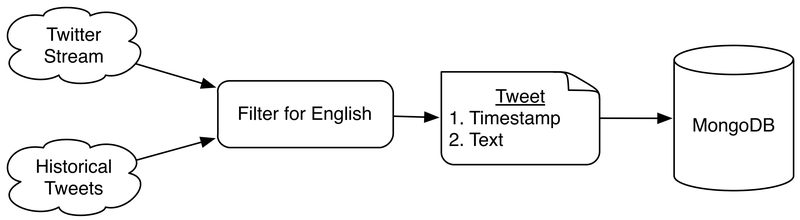

Figure: Data collection process

Sentiment analysis

Sentiment analysis is perform on each tweet. We determine the mood based on Plutchik's wheel of emotions. Plutchik's wheel of emotion consists of 8 basic emotions: *Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger and Anticipation*

The method for determining a tweets mood values is to do a word by word comparison of the words in the tweet to a set of words associated with each mood. If a word in the tweet exists in the list for a given mood, the value of that mood for that tweet is increased. Once every word in the tweet is processed, the values for the moods are normalized.

For example the tweet *"I love to eat cake but I fear I will be angry after and look disgusting"* has the following mood score:

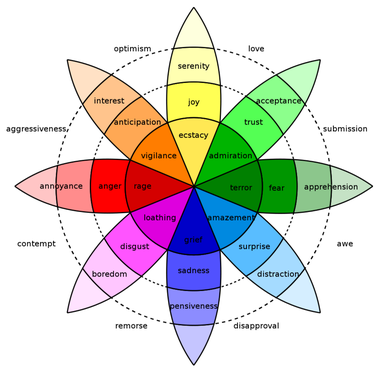

Figure: Plutchik's wheel of emotion.

Joy	Trust	Fear	Surprise	Sadness	*Disgust*	*Anger*	Anticipation
0.0	0.2	0.2	0.0	0.2	0.2	0.2	0.0

To handle the large amount of data, the sentiment analysis is implemented in the Map-Reduce framework Spark. Two months of data already results in approximately 88 million tweets. To process this size of data, we use Spark to calculate the sentiment of the tweets in a scalable fashion. The tweets are streamed into the application for both historical data as well as in periodic batches from the live stream.

Prediction of Stock Prices

We attempt to correlate the daily mood with the closing price of the NASDAQ. The sentiment results from the day are analyzed using a feed forward neural network. Stock prices are aggregated with the average twitter mood data for each day, and this data is used to train the neural network. The last 7 days of stock data and mood data are fed into the neural network to get the stock prediction of the day after. This algorithm has an average of 4.6% error in it's prediction with a maximum error of 9.2%. In the table below are the results of the last 7 days of the test data.

Actual Stock Price	Predicted Stock Price	Percent Difference
42.09	42.45	0.008
40.93	42.25	0.031
40.93	42.23	0.030
40.93	42.51	0.037
38.65	42.60	0.092
39.05	42.24	0.075
39.88	42.00	0.050

Tool and Features

Our tweets are gathered from historical data or from the Twitter live stream and stored into a MongoDB. The Map-Reduce framework Spark is used to process the tweets for the sentiment analysis. Finally, the *ffnet* library in Python was used to build a neural network which we fed the sentiment data from the tweets to.

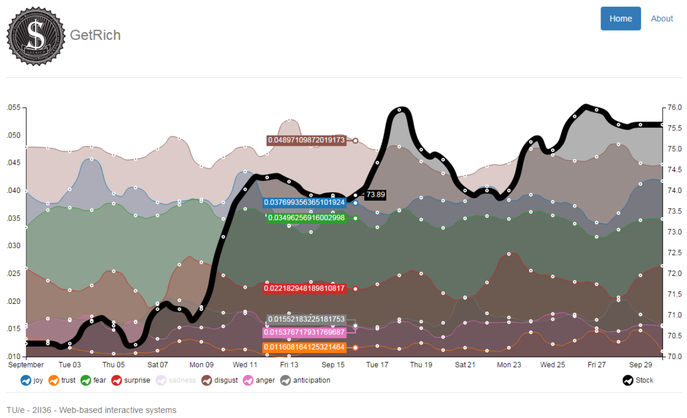

Figure: Website with most recent stock, mood, and prediction data.

Conclusions

We have performed an analysis on a large social network, namely Twitter. We gathered sample tweets which corresponds to about 1% of the total tweets sent, so we will receive approximately 5 million tweets per day. We performed sentiment analysis, which the results are used to create a prediction system for future stock prices. We achieved a scalable solution using the Spark Map-Reduce framework. A prediction of future stock prices is done using a neural network, the results are then stored in a MongoDB. Finally, the results are published on a continual basis on a website. We have implemented all the requirements for running the Twitter analysis with the stock recommender, but we were unable to discover a satisfactory correlation between moods and stock prices with the current configuration.