Data Mining Twitter to Predict Stock Market Movements Summary

Blake Hillier
Advanced Big Data Analysis

February 25, 2020

1 Introduction

This paper is a summary of the paper *Data Mining Twitter to Predict Stock Market Movements* (DMT) by Maxim Pecionchin and Usman Muhammad. Since trading is both a mathematical and emotional activity, adding an emotional component to stock trading models should theoretically improve the model's accuracy. The paper DMT suggests using tweets and ranking them based on 6 moods: happy, sad, excited, calm, dominant, and submissive. They also weight the mood by using pagerank to measure a user's influence on twitter, and then test the lag between sentiment and the stock price. This experiment found certain mood and lag combinations correlates with price fluctuations.

2 Sentiment Analysis

The paper uses the ANEW dictionary, which contains a 3-dimensional rating using valence, arousal, and dominance. Using this dictionary, they were able to create a 6-dimensional label for each stock:

1. Happy: valence > 0
2. Sad: valence < 0
3. Excited: arousal > 0
4. Calm: arousal < 0
5. Dominant: dominance > 0
6. Submissive: dominance < 0

This rating was applied to each word in a tweet, and then each rating was summed up to create a vector describing the tweet. They also wanted to account for the users influence by looking at peoples follower counts. They assume a user with a large amount of mentions while have more influence, which results in a high rank. Similarly, if someone is mentioned only once by a high ranking user, then they must also have a fair amount of influence, resulting in a high rank. However, they ran into a problem with the amount of time and computing power needed to rank this graph. Instead, they use the results of *What is Twitter, a social network or a news media?* by Kwak, H. et al which show this ranking can be accurately estimated by ranking users based on follower count.

3 Experiment

The test was ran over data from July to December 2013. The stock data was hourly closing prices and volume traded of the NASDAQ, and the twitter data was all english tweets with meta data such as follower acount, time, etc. They tested lag times from 0-120 for each sentiment, and found some strong correlations between sentiment, lag time, price, and volume. When happy and calm were lagged 48 and 55 hours respectively, the price was likely to increase, while when sad was lagged 72 hours it was likely to go down. With volume, it was likely to go up when excited was lagged 62 hours and went down when calm was lagged the same number of hours.

4 Conclusion

Twitter data when mapped with the ANEW dictionary and lagged appropriatly can strongly influence the price and volume of the NASDAQ, and could potentially work for other index and even specific stocks, making it an interesting approach worth investigating for sentiment analysis.