Twitter Sentiment Analysis and Stock Market Prediction

Anthony Dionise
MSU Computer Science Student
4341 Appletree Lane
Lansing, MI
dionise5@msu.edu

ABSTRACT

The purpose of this project is to analyze twitter data and determine if it can be used to accurately predict changes in the stock market. I found that it is very difficult to predict changes in the stock market, but it becomes easier as we add variables to our model.

Keywords

Netflix, McDonalds, Facebook

1. INTRODUCTION

With this project, we are trying to accurately predict changes in the stock market based on twitter data. We are analyzing twitter data surrounding a particular stock and determining the polarity of those tweets. If models can be accurately constructed, this method can be used to predict price changes in the market and make a user money. Attempts for this have been made in the past. See for example:

https://arxiv.org/pdf/1010.3003&

http://www.aclweb.org/old_anthology/P/P13/P13-2.pdf#page=72

Predicting changes in the stock market is very difficult. Prices change for an incredibly large number of reasons, and it is impossible to include all of them. Using twitter is just one of many variables that could be included to predict stock market changes. However, Twitter data can be useful if used correctly.

While doing this project, I found that Twitter data can be useful in predicting stock market changes, but it should not be the only resource. The models created using twitter data can predict prices somewhat accurately during some periods of time, and inaccurately during other periods of time. I found that the more variables you add to your models, the more accurate they predict price changes. For instance, if a model is built using price changed from the previous k days and the sum of twitter sentiments from the previous k days, we can add the average twitter sentiment to create a more accurate model. We can take this a step further and add CNBC to the model to make the model even more accurate.

2. PRELIMINARIES

The problem domain for this project is Big Data and Machine Learning. We are trying to see how extremely large amounts of data can be used usefully in the world. With this project, we use large amounts of data to make predictive models that determine changes in the stock market. I have collected and preprocessed vast amounts of twitter data. I also collected historical stock data from Yahoo Finance. The rows include previous day price change, sum of twitter polarity, mean of twitter polarity, and average twitter polarity from the previous day. We are using regression analysts to predict numerical changes in stock prices for specific stocks. We are trying to determine if the twitter mood about a particular stock can help predict a future price change.

3. METHODOLOGY

Twitter data and historical stock price data was used in this project. The first step was to load the twitter data, provided to us on the cse servers. The twitter data is a collection of tweets saved in the form of json. I filtered every tweet object to contain only the date and the content of the tweet. I then used regular expressions on these new tweet objects to collect tweets based on certain keywords. For example, “Netflix” was used to filter out tweets about Netflix. These tweets were then stored in a data frame in Python. By using the sentiment analysis library for Python, I looped through each tweet, determining the polarity. This returns a number between -1 and 1. I also loaded stock data into separate data frames. So, for each stock, I created a linear regression model using the price change from the previous day and the sum of tweet polarities from the previous day. I subsequently created Lasso and Ridge models with the same attributes to try to get better results. I then added the average twitter polarity from the previous day to each model to see if that makes the model more accurate. After this, I added the median of tweet polarity from the previous day. I found that each attribute made the models more accurate.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

1. Many tweets were collected from April 27, 2015 to April 27, 2016 and stored on the cse servers. Fortunately, this data was made accessible to us by our professor. The preprocessed data includes many tweets that are not needed. Each tweet object also contains way too many details about the tweet. Each tweet was filtered down to just the creation date and the text. This is all that was needed from each tweet. The twitter was then filtered down to tweets that are about each stock. For example, tweets about McDonalds.
2. For each model, I calculated R-Squared and Root mean squared to determine the amount of error in prediction.
3. I used Amazon EMR with Spark to filter some twitter data.
4. Hadoop and spark were used to filter large amounts of data.

4.2 Experimental Results

The following contains the experimental code:
https://github.com/tonystonee/TwitterAnalysis

Through this project, I found that twitter is not the best way to predict stock market changes. However, I also found that more

variables create a more accurate model, depending on the variable. I believe Twitter is an excellent source when creating predictive models on stock market price changes, but it should not be the only source.

4.3 Discussion

This biggest thing I learned from this project is that adding variables to predictive models make them more accurate. Twitter data should be used in the future to make predictive models, but perhaps in more creative ways.

5. CONCLUSIONS

This experiment has shown that predictive models perform better when they are trained with a larger amount of variables. In the future, we should try to find more creative ways to use the twitter data, and add other sources of data.

6. REFERENCES

[1] Bowman, B., Debray, S. K., and Peterson, L. L. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst., 15*, 5 (Nov. 1993), 795-825.

[2] Ding, W., and Marchionini, G. *A Study on Video Browsing Strategies.* Technical Report UMIACS-TR-97-40, University of Maryland, College Park, MD, 1997.

[3] Fröhlich, B. and Plate, J. The cubic mouse: a new device for three-dimensional iput. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '00)* (The Hague, The Netherlands, April 1-6, 2000). ACM Press, New York, NY, 2000, 526-531.

[4] Lamport, L. *LaTeX User's Guide and Document Reference Manual.* Addison-Wesley, Reading, MA, 1986.

[5] Sannella, M. J. *Constraint Satisfaction and Debugging for Interactive User Interfaces.* Ph.D. Thesis, University of Washington, Seattle, WA, 1994.