Predicting Stock Trends using Twitter Data

Nicolas Broeking
Department of Computer Science
University of Colorado Boulder
CSCI 5502
nicolas.broeking@colorado.edu

Anna Hoffee
Department of Applied Math
University of Colorado Boulder
CSCI 4502
anna.hoffee@colorado.edu

Joshua Rahm
Department of Computer Science
University of Colorado Boulder
CSCI 5502
joshua.rahm@colorado.edu

ABSTRACT

With social media on the rise, the amount of data available for processing is growing at an increasing rate. It is an advanced area of research to be able to use that data to predict or better understand the world around us. In our project we will attempt to use Twitter to determine if, and how strong of, a relationship exists between public sentiment and chosen stock values.

Categories and Subject Descriptors

H.3.4 [Information Systems Applications-Systems and Software]: Information networks; J.4 [Social and Behavioral Sciences]: Economics

General Terms

Algorithms, Measurement, Economics, Experimentation, Human Factors

Keywords

Social Networks, Trends, Blogging, Tweets, Hashtag, Twitter, Stock

Goal

To determine if there is a relationship between the sentiment of tweets and the change in stock price for the next day. Then to create a user friendly mobile application to display real time stock predictions.

1. INTRODUCTION

The recent rise in popularity of online social media has led to a huge amount of social data available online for analysis. The popular social media site, Twitter, gets an average of 6,000 tweets posted per second and 500 million per day. These tweets can be analyzed to determine future market values. For the purposes of this project we looked at four companies, Amazon, Apple, Google, and Samsung.o

The first step is to open a connection to the twitter streaming api. Using this api we receive anywhere between 1% and 40% of all the tweets being tweeted. This percentage depends on the current Twitter load. We collect the stock information from yahoo finance. We store the stock data for the same time period we get the stock data for multiple granularities.

The next phase in our pipeline is the data storage step. We are using sqlite3 as our database to store all of our tweets. Each tweet stores information related to the user and the tweet to determine possible variables to match against for the statistical analysis.

The third phase is to pre-process the data. Before we process the data we need to filter based on language. We only accept tweets that are English and have no images or videos. Once we have our English tweets we start categorizing the tweets. We separate each tweet into four different bins, one for each company. We then change all URLs to URL and all hash tags to TAG. This allows us to use these as features for the sentiment analysis.

The fourth phase is our sentiment analysis phase. We use a naive Bayesian classifier to detect whether a tweet is positive. Once we have applyed our sentiment analysis to the data send it to our next stage.

We are currently working to implement this phase. We are taking the data and trying to detemine a relationship between the sentiment and a change in the stock price.

Our next step is to determine the realtionship between the variables and then implemnt a mobile application that recomends which company to invest in.

2. AUTHORS

The authors of this proposal are Nicolas Broeking, Anna Hoffee, and Joshua Rahm. Broeking and Rahm are graduate

students at the University of Colorado at Boulder. Hoffee is a undergraduate student at the University of Colorado. Nicolas Broeking has worked on embedded systems and mobile applications for the past four years. He would like to take the results from the project and create a user friendly application that allows a user to easily interact with the data and make financial predictions. Josh Rahm has spent his career working with embedded devices, cellular technologies and billing platforms, and is interested in applying data mining concepts to markets.

3. MOTIVATION

The Stock Market is one of the largest entities in Western and World economies. In a world where a 15% increase is assets per year is massive, even the ability to increase certainty in the market by a few percent is a huge. Companies and individuals can harness this technology make billions and secure investments, producing and saving billions for the economy. While making billions is outside the scope of this project, we think it is possible to significantly increase the accuracy of stock predictions. We know that public image is critical to stock value. If a company's image drastically decreases then we predict that its stock value will change. It is our goal to discover how the company's public image, determined through Twitter, will affect its stock price and then to create an application that recomends stocks for a user.

4. LITERATURE SURVEY

Many similar experiments have been done in the field of data mining. The first and probably the closest to our project is Correlating Financial Time Series with Micro-Blogging. This project looks at 150 random companies stock prices over a six month time period. Then they take all the tweets with specific hashtags related to the company and construct a context graph. In this graph the tweets themselves were nodes and any actions on the tweets were edges. They then use this graph to find relationships with the stock price. This project looked for other things than just stock trends though. They wanted to find relationships between the data and how much a stock will change, and what the values of the stock will be. They determined that the most reliable way to determine the information is by looking at the number of edges in the graph.[1]

Another team, Twitter Mood Predicts the Stock Market, attempted to determine how twitter âĂIJmoodâĂİ effects the stock price. They grouped tweets into 6 dimensions Calm, Alert, Sure, Vital, Kind, and Happy. They then took these categories and created a relationship between to the closing value of the Dow Jones. They found that there is a correlation between what people are posting and how the stock market as a whole performs. [2]

Many other studies have been done similar to the first two on how to use twitter to predict stock prices. The one thing that has been determined for sure is that there is a high correlation between peoples' attitudes and stock value. Twitter was used in 2013 to predict a drop in Royal Caribbean's stock value when people started tweeting about the flu spreading on one of its cruises. Researchers were able to predict the drop in price 48 minutes before the stock plunged about 3%.[3]

5. TASKS

In order to achive this task we split our project up into 7 tasks.

5.1 Data Collection

5.1.1 Stock Collection

Yahoo Finances allows us to gather historical stock data for each of the four companies. This data source provides us with the Open, Close, High, Low, and Data attributes. We use this information to find a correlation between the sentiment of the tweets and the change in stock price.

5.1.2 Twitter Collection

Twitter data is collected from the Twitter Streaming api. This api provides us anywhere from 4% to 40% of all tweets being posted depending on the current load that twitter has to handle. As apart of the data collection phase we filter our tweets. If a tweet contains any reference to any of the compainies we sort it. This allows us to dramatically decrease the amount of tweets that we are going to store. After all of our filtering we still have a massive amount of data. Each week ends up containing about 2GB of data. This then brings the total amount of tweets to over 10GB for about 5 weeks of collecting tweets.

We are storing these tweets in a sqlite3 database. We decided to go with a sqilte3 database because it is fast and portable. We choose a relational database schema because we needed to optimize our next phases for speed. Beause we could have over 20 gb by the end of the semester we need a way to be able to easily and quickly read data from the database. For each tweet, we store: if the tweet was re-tweeted, how many times the tweet was re-tweeted, the date and time, the user id, the tweet id, the users followers count, the users friend count, the user's name and finally the tweets text. We want to store as much infomation as possible because of the time it takes to gather the data we arnt able to collect more if we need more.

5.2 Data Preprocessing

In the preprocessing step we need to filter our tweets on certain criteria. Logically we do everything in the data preprocessing step but we acctually do them in two different phases. The first part of our preprocessing is done during the collection phase. First we filter all tweets based off of if they contain a reference to one of the four companies. As apart of the twitter api we are able to specify that we only

want tweets if they contain the text Apple, Google, Samsung, or Amazon. Once we receive the tweet The next step is to filter based off of language. Because analyzing tweets in multiple languages makes the problem very difficult we only store english tweets. Once we filtered our tweets we store them in the database.

Once we have done completed the data collection phase we do the next step of the data preprocessing. In this step we prepare the data for sentiment analysis. In order to do this we first bin the tweets into four different categories based off of a reference to the company. For example if a tweet contains "Apple" we put into Apples bin.

5.3 Sentiment Analysis

Once we have preprocessed all of the tweets we can finally analyze the text segment. To perform sentiment analysis we use a naive bayes classifier from the nltk python library. We train the classifier on 5000 tweets and then test its accuracy with 10000 tweets. Each one of these tweets was hand classified and publicly published online. We are only interested in the ratio between positive tweets to total tweets so we trained our classifier to mark anything that was not obviously positive as negative. We did not need to include a neutral category.

The major part of the sentiment analysis was choosing features to use for our classifier. We experimented with using different features such as letters, user attributes, and tweet attributes but for the most part these led to very low accuracies. Finally we found that using the words yeiled the best result. In order to determine the features we first take the training data and calculate the probabilties for the naive bayse classifier. Once we have trained the data we take the top 50 infulential features and remove all others from the list. We do this to decrease the run time of the sentiment analysis.

The major challenge with the naive basian classifier is getting a high accuracy. Even with our best features we were only able to get our accuracy higher than 64%. We were not initially satisfied with this accuracy however in our research to create a better twitter sentiment engine we talked with Jamey Wood, CTO of Wayin. Wayin specializes in twitter sentiment analysis and does it on a daily basis for many custimers. He told us that 64% is acctually a very high accuracy for this kind of analysis in industry and so trying to raise this accuracy would be a waiste of time. However we still want to look at other possibilities if time allows that could possible increase the accuracy of our sentiment analysis.

The last major challenge of the sentement analysis is the runtime. We have a huge trade off with the number of features, types of features and how long it takes to run. Currently, after we limit our features it takes about a second to classifiy a tweet. To classify a weeks worth of tweets it takes almost 72 hours. Anytime we train our classifier or add more features it dramatically increases the time it takes to classify the tweets.

5.4 Correlation Analysis

Now that the sentiment analysis is completed and we have all the tweets classified as positive or negative we can begin the correleation analysis. We combined the twitter data with historical stock data. The stock data is obtained from http://www.nasdaq.com/quotes/. Specefically we are looking at Open and Close values. Before starting correlation analysis the twitter data and stock had to be combined. This was done all in R. The next step is to regress the change in open and close values for a given day on the percent of positive tweets about the company from the previous day. To be able to do this the twitter and stock data need to be matched up correctly. Because there are days of missing twitter data that cannot be recovered and no stock data on weekends there was some pre-processing involved in matching up tweets from one day with open and close values for the next day. Now we have an R script that is generalized to any set of stock and twitter data with arbitrary patterns of missing days, so processing of further data will be quicker. Currently there aren't enough data points to come to a good conclusion about, but soon there will be a lot more.

Looking ahead, after more data is collected the regression equation will be as follows: $Y = \beta_0 + \beta_1 X$ where Y is the change in open and close values, and X is the percent of positive tweets from the previous day. We are anticipating to see positive correlation between the percent of positive tweets and the change in open and close values. To test this correlation we'll use the chi-squared test for independence, the lift calculation, and the t-test to test the hypothesis that $\beta_1 = 0$ indicating that X is independent of Y. To fit the regression model we will use the ordinary least squares estimator. We're performing this analyis on data from Google, Samsung, Apple, and Amazon. As well I anticipate that the stock data/twitter data for each of those companies will not be independent of each other, so they could potentially be added as regressors in each others models. Once more data is collected we can begin the statistical analysis.

5.5 Error Analysis

To evaluate the accuracy of the model we'll look at the R^2 value, which shows the percent change in Y that is explained by X. We can also look at the standard errors the of regression coefficients and the residuals of the models. If there is lots of correlation in the errors we can try using the generalized least squares estimators instead.

5.6 Application

Our goal is to create a mobile application that can process tweets in real time and then display what companies it recomends investing in. In order to acomplish this task. We will take our model that we developed using the stages above

and using a server to analyze tweets in real time. The server will then use the model to predict the stock values for the next day. When the user opens the app on his phone the app will reach out to the server using a web request and get the recomended stocks using json. The mobile device will then display these results to the user.

5.7 Sources

JOSH CAN YOU USE BIBTEX