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 a data warehouse 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 know 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.

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

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. These tweets are stored in a sqlite3 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..

5.2 Data Preprocessing

In the preprocessing step we need to filter our tweets on certain criteria. The first thing we do is filter all tweets based off of if they contain a reference to one of the four companies. For simplicity if a tweet contains any of the words, Apple, Google, Samsung, or Amazon we store it. 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. The final step in pre processing the data is to change all URLs in the tweet to URL and to change all hashtags to TAG. This allows us to create a feature that we can use to do sentiment analysis in the next step.

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. We train the classifier on 5000 tweets and then test its accuracy with 10000 tweets. 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.

Performing sentiment analysis on tweets proved to be a much harder task than initially thought. 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 taking the 50 most positive words from that training set and using these as features for the classifier yielded the highest accuracy.

The accuracy of our classifier is about 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. 64% is acctually a very high accuracy for this kind of analysis in industry and so increasing the accuracy much higher seems to be an impossible task.