Using Textual Data to Predict Stock Fluctuations within the Technology Industry

Project Proposal

CSCI 5622

Mary Letey

Morgan Allen

Colton Williams

Aniq Shahid

In order to predict notable increases and decreases in the spot prices of key companies in the technology industry, we will focus on modeling textual data, such as articles from reputable financial news sources and legal filings, against historical stock price fluctuations. A company's stock price depends heavily on the investors' perception of the company, such as perceived growth and perceived risk. Many financial news sources can release information that influence this perception. In addition, legal filings (such as SEC filings) can contain pertinent information regarding changes in the company, which effect stock prices. Using corporate filings to find useful information on shifts within a company is a strategy for leading executives at Berkshire, and common practice when performing financial analysis on a company to predict future performance.

However, given the amount of data contained in articles and corporate filings, it's very laborious and inefficient for financial analysts to sort through this information directly. Enter machine learning! Using textual analysis modeled against historical stock prices, our team hopes to find key-word predictors for fluctuations (negative or positive) in companies across a specific industry. We are focusing on a single industry because we speculate that the predictors will be correllated among similar companies.

To find when the fluctuations occur, we will be using numerical data on the spot price. All numerical data will be available via Bloomberg (Mary Letey has access through the Leeds school of business). The textual data will be obtained from news sources, such as the Wall Street Journal and Financial Times, as well as SEC filings. This data is freely available, but we will have to download and assemble it manually. Furthermore, we can use Google Trends to obtain correlation between certain terms to learn more about the state of company. For example, checking how often terms such as "layoffs", "business expansion", and "product release" are searched for a given company may have a relation with its stock price.

In order that the textual data isn't the only variable included in our model, we will include relevant numerical variables in our analysis, such as spot, options, futures, indexes, etc. We will also use Google Trends keywords to provide additional sentiment analysis. The industry will be modeled using a weighted mix of leading companies in that industry.

We have not finalized the training/learning model we will be using for stock prediction, but based on our research and knowledge, neural networks and Support Vector Machines (SVM) are good candidates. Time permitting, we will also attempt to develop a hybrid model wherein we can use a combination of learning/training algorithms to achieve better overall accuracy.

The exact stock prediction window is also unknown at this point, but we will attempt to make short term predictions (1 day, 3 days, 1 week) at first, and then expand the model to make longer term predictions.