Project Proposal
Event-Driven Stock Prediction using NLP

Anant Jain, Ilya Yudkovich
CS6120 NLP Spring 2019
jain.anan@husky.neu.edu,
yudkovich.i@husky.neu.edu

Introduction

It has been stated that the price of a security reflects all of the information available, and that everyone has a certain degree of access to the information [Efficient Market Hypothesis (EMH), Fama 1965]. Today, a lot of new information related to companies and corporations listed on various stock exchanges like NASDAQ, NYSE and AMEX appear constantly on the web, making an instant impact on the companies' stock value. Surveilling all this information in real time is crucial for big trading firms but is not very much in reach of a lone investor. In this project, we present a news surveillance system targeting financial news to predict its impact on some of the top market cap companies. We aim to tackle the challenge from the position of a single investor without any access to real time trading infrastructure.

Related Work

Despite years of research and studies, there are still debates about what information could best predict the volatility of stock market. In the field of Artificial Intelligence, there have been previous attempts in designing models based on historical and time series data [Taylor and Xu, 1997; Andersen and Bollerslev, 1997; Taylor, 2007]. However, these methods do not account for one of the key sources of market volatility which can have dramatic effect on a security's share i.e. financial news [Cutleret al., 1998; Tetlocket al., 2008; Luss and d'Aspremont, 2012; Xieet al., 2013; Wang and Hua, 2014].

Recent improvement in computing power and introduction of new NLP techniques enabled to address this issue and make use of financial news. Since then, there have been many attempts to boost the accuracy of market predictions using language features, such as, tree representations of information [Xie et al., 2013], identification of expert investors [Bar-Haim et al., 2011] and risk based on financial reports [Koganet al. 2009]. The paper by Engelberg (2008) discussed how linguistic information in the text has a greater long-term predictability for prices than using term frequencies. Although very useful, these techniques do not account for structured relations in text which limits their potential.

Structured representations can be found using open information extraction (Open IE) tools and take semantics in account, but that leads to increased sparsity which in turn limits the predictive power.

To this end, this project focuses on learning event embeddings. Event embedding are dense vector matrices which are trained such that similar events would have similar matrices even if they don't comprise of common words.

Apart from that, sentiment analysis is another way of doing deep semantic analysis of news articles [Das and Chen, 2007; Tetlock, 2007; Tetlocket al., 2008; Bollenet al., 2011; Siet al., 2013]. Their work is more or less orthogonal to what we plan to achieve.

Dataset

We plan to create a large-scale financial news dataset by crawling Reuters.com. It is one the biggest international news organizations along with Bloomberg and specializes in international business and financial news.

For the preliminary crawls, we decided to crawl news for the past 1000 days for the top 37% market-cap companies (~2590 tickers) and store the data in a highly compressed pickle of a pandas dataframe. As expected, higher market-cap companies have more news as compared to lower market-cap companies, but since we are not concerned with the companies themselves but the type of events they undergo, we can safely discard ticker information from the data.

To build the labels for each event. We plan to crawl historical stock prices (Open, Close, Adj Close, Date) for each ticker from Yahoo Finance and calculate their relative return w.r.t. S&P. Depending on the relative return value, we plan to segregate the returns into clusters (Strong Sell, Sell, Buy, Strong Buy) using quartile ranges.

We plan to treat history news as daily event sequences and are currently debating to build more than one label: short (1 day), mid (7 days) and long (28 days), which correspond to relative return over the length of investment. Xie et al.[2013], Tetlock et al.[2008] and Ding et al.[2014] show that the performance of daily prediction is better than weekly and monthly prediction. But, despite of relatively weaker effects of long-term events, the volatility of stock markets is still affected by them.

After merging the news and the labels by date, at the end, our dataset would comprise of a **news** column containing the news about an event, and a **label** column indicating what would have been the best move to make given the news.

Methodology

As evident it is a supervised machine learning text classification problem.

Given the news, we first plan to clean the news corpus, normalize the text, remove stop words and filter it such that the corpus would contain only those words which appear frequently in financial articles.

Next, in order to process the embeddings on a laptop, we limit the maximum number of words per event to something our laptops can handle and pad the sequences. We plan to use glove[1] embeddings for words.

For the predictive model, we plan to use deep learning in order to capture the influence of news events over a history. Usually, when using embeddings, density estimators (convolution neural network) can achieve good results but can misbehave in high dimensions [Bengioet al., 2005]. Therefore, we first plan to use a 1D convolutional neural network (CNN) to perform semantic composition over the input event sequence accompanied by a pooling layer in order to extract the most representative global features and associate them with stock trends through multiple shared hidden layers and an output layer.

If that doesn’t work, we plan to try a novel neural tensor network (NTN), which is capable of learning semantic compositionality over the arguments of an event by combining them multiplicatively instead of only implicitly, as with standard neural networks.

Evaluation

We can use the standard evaluation procedure in machine learning i.e. making train-validation-test splits of our final dataset, training on the training data, tuning on validation data and testing on testing data. We plan to use accuracy and precision as our metrics, as our labels are going to be balanced because of segregation of relative return by quartiles.

[1] http://www-nlp.stanford.edu/pubs/glove.pdf