Predicting the Stock Market with NLP

Author	V-Number
Tim Chan	803066
Adam Dubicki	799637
Shengzong He	785261
Lucas Sherman	711371
Haodong Tau	836169

Abstract

ABSTRACT GOES HERE

1 Introduction

The goal of this project is to use news headline data sets to predict the rise and fall of the United States industrial stock market.

To achieve this, labeled news headlines and supervised learning techniques will be used to create an accurate stock market classifier. This project is interesting for several reasons. Most of the data set pertains to international headlines. If a classifier can be created from it than it will illustrate how news from around the world affects the US. It will also reveal if popular aggregate news suffices to identify trends in the stock market.

Stock market prediction is a difficult challenge due to the many outside factors which affect it. Natural disasters, scandals, and product hype can all contribute to people buying or selling stock. Many of these factors can be captured by news aggregation. If the classifier built is sufficiently accurate than it could be extended to make predictions off of additional news aggregation sources.

1.1 Data Description

The data set for our chosen problem comes from Kaggle: a website which hosts various data mining and machine learning competitions. It was posted in August 2016 by user Aaron7sun (”Daily News for Stock Market Prediction”, 2016). Reddit.com is a news aggregation website where users can vote on news headlines. The news headlines in the data set are from the world news section. Reddit.com has a large amount of US visitors, resulting in many US specific headlines reaching the top 25 posts daily, despite the focus of the World News section being international headlines.

The data is packaged as a 6MB CSV each row containing news headlines and binary class. There are three files, RedditNews.csv, DJIA_table.csv and Combined_News_DJIA.csv. RedditNews.csv contains top 25 news headlines for a given day decided by the number of votes by . DJIA_table.csv contains the stock data from Yahoo finance, and is based on the Dow Jones Industrial Average. (”DJI Historical Prices - Dow Jones Industrial Average Stock”, 2017).

The Dow Jones Industrial Average (DJIA) is an indicator of the current state of the American industrial sector. It indicates how 30 large publicly owned companies based in the US (such as Microsoft, Verizon, Visa, IBM) have traded during a standard trading session in the stock market.

The Combined_News_DJIA.csv file gives the top news headlines for every given day along with label entry saying whether the DJIA rose/stayed or fell. A row contains the Date, DJIA label, and then the top 25 headlines ordered with the number of votes. The combined file has 1990 rows, each corresponding to a day. Most of our working was performed on Combined_News_DJIA.csv file.

In terms of data distribution the dataset is slightly biased towards the DJIA rising (52% of the data is of class 1). The baseline to improve upon is 52%, because naively guessing 1 for every classification would yield this accuracy.

1.2 Related Work

There are many ways that Natural Language Processing could assist in the movement of the stock market. For example, using the AlchemyData News API provided by IBM could allow machines to gather real-time intelligence on specific stocks, understand and respond to news events, monitor company sentiment and etc. (Sedlak, 2016).

There are also researchers that used NLP and conducted a Textual Analysis for stock price prediction. The study lead by Stanford University introduced a system that forecasts company's stock price changes and classifies them into UP, DOWN and STAY. By building a well defined corpus that aligns the descriptions of financial events, the system's Unigram model could predict the stock trend with an accuracy of 54.4%. The low accuracy is because that newspaper articles reflect not only new information about the company, but also the perspective and options of third parties, which may require more time for the market to digest (Lee et al, 2014).

Two researchers one from Iona College and the other from University of Arizona worked together and performed textual analysis of the predicting the stock market using recent financial news. The researchers approached the problem using prediction with machine learning using financial news articles analysis where they researched with different textual representations, such as bag of words, noun phrases, and named entities (Schumaker, R. P., & Chen, H. 2009).

The researchers investigated 9,211 financial news articles and over a five week period of 10,259,042 stock quotes that covered 500 of S&P stocks. They started their analyses of a stock price twenty minutes after a news article was released. A support vector machine (SVM) was used for discrete numeric prediction and models containing different stock-specific variables. They showed that the model that had both article terms and stock price at the time of the of the articles release was the closest prediction to the actual stock price. They achieved a prediction of the price moving in the same direction as the future price with a 57.1% accuracy and their highest return was 2.06% which used a simulated trading engine. They also investigated different textual representations and found that Proper noun scheme performs better the bag of words (Schumaker, R. P., & Chen, H. 2009).

One report used sentiment analysis on Twitter posts to try and predict the daily closing values of the Dow Jones. The focus was on finding out whether the mood of the society was correlated to the stock price of the Dow Jones. Two different sentiment analysis tools were used: OpinionFinder and Google Profile of Mood States. OpinionFinder analyzed only a positive and negative moods, while Google Profile of Mood States would measure mood in term of 6 different dimensions. Correlation between mood and stock price was determined using Granger causality analysis. A Self-Organizing Fuzzy Neural Network was used to do the actual classification, which achieved an accuracy of 86.7%. Different moods were found to give different classification accuracies (Bollen et. al., 2010).

Researchers from the Malaysia has performed a study on text mining of news-headlines targeting specifically to the Foreign exchange (FOREX) market prediction. The study resulted the development of a multi-layer dimension reduction algorithm that consisted of three main layers: Semantic Abstraction Layer, Sentiment Integration Layer and Synchronous Targeted Feature Reduction. The proposed algorithm resulted with a model that has accuracy of up to 83.33% (Nassirtoussi et. al. 2015).

It is worth mentioning that in the semantic abstraction layer, the researchers addressed the problem of co-reference in text mining. Co-reference occurs in the situation where two or more words in a text corpus is referring to the same entity. The problem with co-reference in text mining is that it increases the dimension of the mathematical model and decreases the performance of it. By using customized approach of extracting heuristic hypernyms and feature selection, the semantic abstraction layer resolved the issue created by co-reference in a large text corpus. It reduced the unnecessary dimensions in the training model caused co-reference words and increased both the performance and accuracy of the algorithm.

2 Methods

Below details the methods applied to try and yield the highest accuracy for predicting the stock market. Initially, a baseline was implemented based off a tutorial's code. Afterwards, the baseline was extended based on feedback from other peers.

2.1 Baseline

As part of the Kaggle competition, various tutorials are posted to help get users started with processing the dataset, The highest voted tutorial was provided by Kaggle user Andrew Gélé. In the tutorial scikit-learns vectorizer is applied to transform the headlines into word vectors. The vectors contain the frequency of bi-grams in the headlines. The vectors are then processed and fitted with logistic regression. This approach yielded an accuracy of 56.3% (Gélé, 2016).

For Gélés approach, the dataset was split by date such that all headlines from 2015 onward were test data. All other headline instances are part of the training set. This split was also used for our experiments for two reasons.

First, it aims to minimize multi-day headlines from being part of the training and test set. For example a week long event such as the Olympics could be highly correlated with the DJIA rising. If the data set is shuffled than there is a high probability of these headlines could be split across the training and test set. The classifier would essentially be training and testing on practically the same headlines, generating an optimistic accuracy. By splitting by date, related events shall be mostly segregated into the training and test sets. The only exception to this shall be events at the boundary, which will unavoidably be split among the two sets.

Secondly, it is a very common split among other Kaggle competitors. By using the same split, our results are comparable to the results of other competitors.

To begin the experiments, a testing framework was built off the tutorial code in combination with other scikit learn libraries. The experiment setup in python aimed to create a highly iterable framework. The python file created was a preprocessor which would vectorize the data. The preprocessor also performed any data preparation steps such as stemming, stop-word removal and day offsetting.

Day offsetting is a minor improvement experimented with where the data is shifted forward or backward a certain amount of days. The unshifted data has headlines paired with stock results of that current day. This assumes that the headlines have made an impact on the current day. By offsetting, headlines for day i are paired with the binary class for day $i \pm o$, where o is the offset. The purpose behind this feature is two-fold. Examining positive shifts assumes that headlines have a latency towards affecting the a market. Negative shifts show how insider information and news leaks might allow for predictive trading.

Next a generic processor class was created. The processor takes as input a data set from the preprocessor, and a classifier from the python library scikit-learn. The processor splits the data into training and test sets, and trains with the training set. Finally, it forms predictions on the test set and outputs various metrics such as f-measure, accuracy, precision and recall.

After experimenting with several classifiers, the three best performing classifiers were logistic regression, SVM (linear kernal), and stochastic gradient descent. Through experimentation it was also determined that the a mix of bigrams and trigrams often yielded the highest results.

2.2 Extending the Baseline

After receiving feedback, the baseline method was extended in an aim to improve the accuracy.

- The world news data set does not always contain information relevant to the stock market. In an attempt to get more pertinent data, other subreddits (Economics, stocks, news (USA), and technology) were scraped using an API bot.

- The offset is an interesting feature which got minimal experimentation in the baseline. Additional offsets were attempted to see if future news insight had an affect on the market.

- While the headline date was used to split the data, it was never introduced as a feature. It

was determined that the date could be pertinent if there were fluctuations around certain seasons or holidays.

- Sometimes world events will span across multiple days of headlines. As discussed earlier, shuffling the data can cause cross contamination of headlines between the train and test data. Yet, it still may be important to reflect that events are trending. To represent this as a feature, a relation attribute was added to capture the similarity of a headline i to its surrounding headlines. It was calculated by taking the sum of cosine similarities between the headlines of day i with days $i+1, i+2, i-1$, and $i-2$. From observation this usually value ranged between 0 and 2. Headlines on the boundary (such as day 1) used a cosine similarity of 0 when calculating days outside of the valid range.

A problem with the reddit is that several of the subreddits were newer, or had less than 25 posts per day. This resulted in the new data sets having slightly different distributions and sizes.

Table 1. Data Set Info

Source	Rows	Columns	% Class = 1
/r/Economics	1981	10	cell6
/r/stocks	667	10	cell6
/r/Technology	1985	15	cell6
/r/news(US)	1984	15	cell6
NY Times	1981	25	cell6

Where /r/ means the data was scraped from Reddit's API. Rows is the number of days, and columns is number of headlines per day. For the NY Times data set, the article headlines did not have a set rank for popularity. To generate a simple heuristic order, the headlines are sorted by length descending.

3 Results

Many of the tables share common feature names. For brevity some feature names have been abbreviated as follows.

LR	Logistic Regression
SVM	Support Vector Machine (Linear Kernel)
SGD	Stochastic Gradient Descent

References

Daily News for Stock Market Prediction. Kaggle.com. Retrieved 14 May 2017, from https://www.kaggle.com/aaron7sun/stocknews

Bollen, Johan & Mao, Huina & Zeng, Xiao-Jun. (2010). *Twitter Mood Predicts the Stock Market.* Journal of Computational Science. 2. . 10.1016/j.jocs.2010.12.007. Retrieved 22 July 2017, from https://arxiv.org/pdf/1010.3003.pdf

DJI Historical Prices — Dow Jones Industrial Average Stock - Yahoo Finance. (2017). Finance.yahoo.com. Retrieved 14 May 2017, from https://finance.yahoo.com/quote/%5EDJI/history?p=%5EDJI

Gélé, Andrew. (2016, November). *OMG! NLP with the DJIA and Reddit! — Kaggle.* Retrieved 14 May 2017, from https://www.kaggle.com/ndrewgele/omg-nlp-with-the-djia-and-reddit

Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. (2014, May). *On the Importance of Text Analysis for Stock Price Prediction.* In LREC (pp. 1170-1175). Retrieved 14 May 2017, from http://www.lrec-conf.org/proceedings/lrec2014/pdf/1065_Paper.pdf

Nassirtoussi A.K., Aghabozorgi S., Wah T.Y., Ngo D.L.C. (2015). *Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment.* Expert Systems with Applications, Volume 42, Issue 1, 2015, Pages 306-324, ISSN

0957-4174. Retrieved 22 July 2017, from http://www.sciencedirect.com/science/article/pii/S0957417414004801

Schumaker, R. P., & Chen, H. (2009). *Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS), 27(2), 12.* Retrieved July 15 2017, from http://dl.acm.org/citation.cfm?id=1462204

Sedlak, M. (2016). *How Natural Language Processing is transforming the financial industry - IBM Watson.* IBM. Retrieved 14 May 2017, from https://www.ibm.com/blogs/watson/2016/06/natural-language-processing-\transforming-financial-industry-2/

Trastour, S., Genin, M., & Morlot, A. (2016).*Prediction of the crude oil price thanks to natural language processing applied to newspapers.* Retrieved 14 May 2017, from http://cs229.stanford.edu/proj2016/report/GeninMorlot\Trastour-PredictionOfTheCrudeOilPrice\ThanksToNLPAppliedToNewspapers-report.pdf