Using Sentiment of News Articles to Predict Stock Price Performance.

David Rodriguez-Romero, Dr. Debzani Deb
Department of Computer Science
Winston-Salem State University, Winston-Salem, NC

Purpose

- **Motivation**: Stock prices are determined by the human investors who respond to publicly available information, therefore, published news articles can effectively influence the stock price values.
- **Goal**: The goal in this study is to use data science in finding a correlation between news articles and their predictive power, if any, on stock prices.
- **Hypothesis**: We hypothesized that the sentiments of the news articles, which were published in recent past, effectively can predict the stock prices in immediate future.

References

1.Reuters, https://www.reuters.com/
2.Nasdaq, https://www.nasdaq.com/
3.Yahoo Finance, https://finance.yahoo.com/
4.Vader Sentiment Analysis, https://github.com/cjhutto/vaderSentiment
5.Weka 3: Data Mining Software, https://www.cs.waikato.ac.nz/~ml/weka/

Acknowledgement

NSF RIA Award #1600864

Feedback @ http://ibigcloud.altl.org

Input Data set & Methodology

- To create the news articles dataset, we used a web scraping program that collects news articles related to the two prominent financial organizations such as Apple and Facebook from the Reuters [1] and Nasdaq [2] websites during the period of November 2018 to February 2019.
- The relevant stock dataset is created by downloading the AAPL and FB stock prices from Yahoo Finance [3].
- We organized both datasets according to the publication dates and joined them. Our resultant data set contains the attributes such as publication date, headline of the corresponding article, actual text content of the article, the financial organization that the article is referring to (AAPL or FB), and % daily gain of the stock price on that day.
- Headline and text then further processed to be tokenized, to remove all stop words, and to perform stemming. We then ran Vader sentiment analysis algorithm [4] that focuses on word and adjective polarities of a word and produces a scores to four classes of sentiments such as negative, positive, neutral and compound (aggregated score).
- In our first experiment, we compare the polarity of the compound sentiment score to the polarity of the % daily gain. When both of them are in the same polarity, we classify that as an correct correlation. Accuracy is measured by % of correctly correlated articles.
- To understand the time lag between when the news article is released and when the market has moved to reflect this information, we also examined the influence after different durations such as the same day and next day.
- We also used machine learning (ML) and utilized WEKA [5], an open source data mining tool, for predicting stock price polarity based on the sentiment values of the news articles.
- **Classification**: Machine learning algorithms such as Naïve Bayes (NB), K-nearest neighbors (KNN), Decision Tree (DT), and Random Forest (RF) from WEKA were utilized to classify.

Results & Discussions

	Same Day % Accuracy	Next Day % Accuracy
AAPL Articles (607)	54.5%	52.1%
FB-Articles (387 articles)	53.7%	43%

	AAPL Articles % Accuracy	FB Articles % Accuracy
NB	56.8%	57.9%
KNN	69.7%	65.6%
DT	59%	64.9%
RF	70.5%	66.4%

- Results in the first table shows that directly correlating the sentiment of the news articles with the stock daily gain is not an effective way to make prediction.
- For both of our data sets, the prediction accuracy are reduced after a day.
- The second table shows the accuracy that is achieved for various ML algorithms.
- In all experiments, 10 fold cross validation was utilized and accuracy of the results are recorded as % of correctly correlated articles with the daily gain value.
- It is evident from the table that ML approach works better than the direct correlation and Random Forest algorithm tend to perform the best.

WINSTON-SALEM STATE UNIVERSITY