Stock Market Index Prediction using Public Sentiment Derived from Online News

Nicholas Wright

Introduction

Stock prediction has long been a way to gain an edge and increase wealth, with many people trading on hearsay or rumors while others invest hundreds of millions of dollars[1] to gain milliseconds on other traders. This indicates that in trading, information is an incredibly valuable commodity. As milliseconds are more than enough time for a computer to accomplish a task, traders will use any edge they can get to gain actionable information. One commonly used tool is the overall mood of the general public. The Thomson Reuters Company, a financial services provider, has long been an industry leader in providing cutting edge information to traders and in 2013 provided the United States Consumer Confidence Index (CCI) early as part of a subscription service[2]. This index represents how optimistic or pessimistic the United States population feels about the economy in a particular month, this is accomplished through a survey that asks users a series of questions about their spending habits[3]. Days after this announcement was made, the New York Attorney General ruled that the two second advantage [provided to subscribers] is more than enough time for traders to take unfair advantage of their early access to this information as they execute enormous volumes of trades in the blink of an eye[4]. This shows that knowledge about public sentiment can provide a tangible competitive advantage to traders and is a stream of information that can assist in decision making.

Streams of information have changed since the CCI was invented in 1967 and public opinion about economic health can now be measured in a variety of different ways. An increasingly popular way of determining the effect public sentiment has on financial markets is the analysis tweets[5]. This is a popular and strong indicator of sentiment as topics can be filtered based on 'hashtags' and opinions can be collected from a wide selection of strangers. However, it was found in a study done by Bloomberg Professional Services[6] that sentiment expressed through the News led to a greater increase in open to close return than the same sentiment expressed through Twitter.

In order to measure this sentiment a large dataset of news headlines has been collected as well as daily historical performance for five individual market indexes. Sentiment classification

---

[1] Williams, Christopher. "The $300m Cable That Will Save Traders Milliseconds." The Telegraph. Telegraph Media Group, September 11, 2011. https://www.telegraph.co.uk/technology/news/8753784/The-300m-cable-that-will-save-traders-milliseconds.html.

[2] Javers, Eamon. "Thomson Reuters Gives Elite Clients Early Edge." CNBC. CNBC, June 12, 2013. https://www.cnbc.com/id/100809395.

[3] Yousuf, Hibah "Thomson Reuters Ends Privileged Access to Consumer Sentiment Data." CNNMoney. Cable News Network. Accessed February 24, 2020. https://money.cnn.com/2013/07/08/investing/thomson-reuters-consumer-sentiment/index.html.

[4] "Leading Indicators - Consumer Confidence Index (CCI) - OECD Data." theOECD. Accessed February 24, 2020. https://data.oecd.org/leadind/consumer-confidence-index-cci.htm.

[5] Ranco, Gabriele, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. "The Effects of Twitter Sentiment on Stock Price Returns." *Plos One* 10, no. 9 (September 21, 2015). https://doi.org/10.1371/journal.pone.0138441.

[6] Cui, Xin, Daniel Lam, and Arun Verma. "Embedded Value in Bloomberg News & Social Sentiment Data." *Bloomberg Professional Services*, 2019. https://www.bloomberg.com/professional/sentiment-analysis-white-papers/.

will be divided into positive, neutral and negative sentiment. With positive sentiment indicating an increase of value in the market index, negative indicating a decrease and neutral accounting for the periods where very little fluctuation is seen in the markets. The research questions this project will address is, Does a significant correlation exist between public sentiment and the performance of major North American Stock Markets.

Literature Review

Stock market volatility has long been a factor when it comes to market health, with certain indexes, such as the Dow Jones Industrial Average being regarded as more stable than others. This has led people to consider how public perception affects certain markets. In a study done by Audrino et al[7], it was found that when using a predictive regression sentiment and attention variables have predictive power for the future market volatility. This study also utilized a heterogenous autoregressive model (HAR) model[8] when forecasting volatility over a one day and weekly period.

Determining a threshold for sentiment classification is key as it provides the basis on which our predictive model will be trained. In the paper "Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis" by Khedr et al[9] an increase of 3.59% was seen in the Naïve Bayes Model through the addition of news polarities and historical stock prices. This indicates that a model can be improved upon through the addition of sentiment-based data.

In the paper "A novel stock evaluation index based on public opinion analysis" by Yin et al[10] a Natural Language Processing (NLP) model was created then finance data from the RESSET database used to analyze the data through correlation, time series and regression analysis. It was found that stock price fluctuation is more sensitive on a day to day scale, no correlation existed between individual investors and finance consultants and the change in Stock market price is represented in public sentiment.

The next paper by Malagrino et al[11] utilizes a Bayesian network to predict how global markets affect the Sao Paulo stock exchange. This is accomplished through comparing market open and close times along with the open and close market price. The different markets were compared over 24 and 48 hour periods and a Bayesian model was trained. No significant correlation was discovered, and the paper suggests moving away from a Bayesian model suggesting that Neural Networks will provide better overall accuracy.

---

[7] Audrino Francesco., Fabio Sigrist, Daniele Ballinari. The impact of sentiment and attention measures on stock market volatility, International Journal of Forecasting, 2019, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2019.05.010

[8] Fulvio Corsi & Roberto Renò (2012) Discrete-Time Volatility Forecasting With Persistent Leverage Effect and the Link With Continuous-Time Volatility Modeling, Journal of Business & Economic Statistics, 30:3, 368-380, DOI: 10.1080/07350015.2012.663261

[9] Khedr, Ayman E., and Nagwa Yaseen. "Predicting stock market behavior using data mining technique and news sentiment analysis." *International Journal of Intelligent Systems and Applications* 9, no. 7 (2017): 22.

[10] Yin Ni, Zeyu Su, Weiran Wang, Yuhang Ying, A novel stock evaluation index based on public opinion analysis, Procedia Computer Science, Volume 147, 2019, Pages 581-587, ISSN 1877-0509. https://doi.org/10.1016/j.procs.2019.01.212.

[11] Malagrino Luciana S., Norton T. Roman, Ana M. Monteiro, Forecasting stock market index daily direction: A Bayesian Network approach, Expert Systems with Applications, Volume 105, 2018, Pages 11-22, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.03.039.

In "Sentiment analysis and machine learning in finance: a comparison of methods and models on one million message" by Renault[12] multiple machine learning methods are used on the same dataset and their performances measured against one another. The methods used were Naïve Bayes, SVM, logistic regression, random forest and Multilayer Perceptron classifier. It was discovered that there was no significant difference between the more complex machine learning models (i.e. Random Forest) and similar methods such as Naïve Bayes. The most significant discovery is that the addition of emojis and punctuation greatly increased the accuracy of the model. For the purposes of this project, punctuation may still be removed as the more formal structure of news articles differs greatly from the social media dataset used in this study.

In regard to the analysis of sentiment in the news, the Bloomberg Professional Services published white paper on Embedded Value of News and Social Sentiment utilizes a machine learning model which to classifies two sentiment scales. The first is a 'story level' sentiment measured from 0 to 100 and a second company level sentiment measured from -1 to 1[13]. For the purposes of this project the -1:1 model will be used to indicate sentiment and all news article will be treated as 'company level'. The methodologies used in this paper to conduct the analysis of news sentiment were also used on Twitter sentiments, it is then inferred that all methodologies discussed above can be utilized and built upon in the dataset and methodology outlined below.

Dataset

The first dataset used in this paper consists of 143,000+ News articles from prominent online news sources, the repository of data was initially downloaded from the competition website Kaggle[14]. The dataset was collected largely between the years 2015 to 2017, however, outliers do exist. The dataset consists of 8 attributes, they are id, title, publication, date, year, month, author, url and content[15]. The publications featured in the dataset are Breitbart, New York Post, NPR, Washington Post, Reuters, New York Times, The Guardian, CNN, National Review, Atlantic, Vox, Business Insider, Buzzfeed News and Fox News. The attributes id and url will be removed and publications will be weighted based on Twitter follower data to ensure an accurate segmentation of public opinion. The second dataset used has been compiled from Yahoo Finance historical data, from the dates of January 2, 2015 to December 29, 2017. The dataset consists of five attributes: Date, Open, High, Low, Close, Adjusted Close and Volume. Five stock market indexes csv's were merged together and sorted by date. The indexes featured are the Nasdaq[16], S&P 500[17], NYSE [18], Dow Jones[19] and TSX[20]. All Attributes from this dataset will be used.

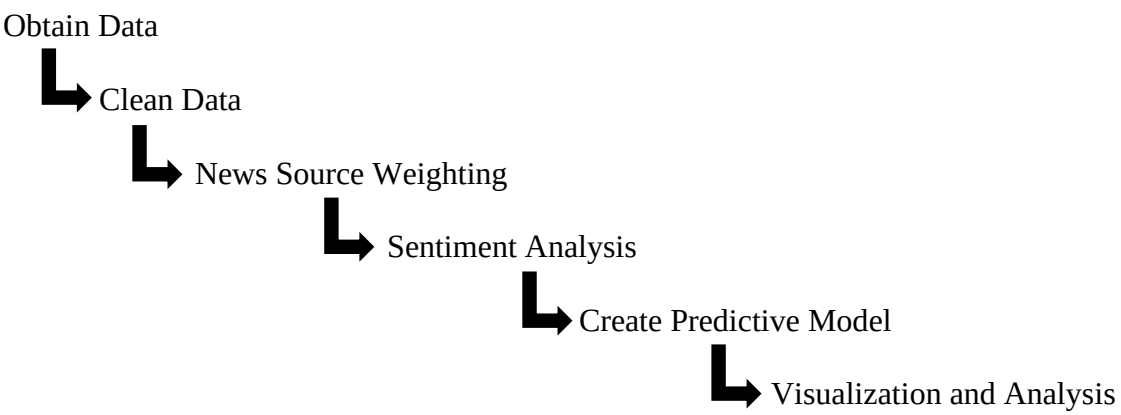
News Headlines		Stock Market Data	
Title	Date	Date	String

[12] Renault, T. Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digit Finance* (2019). https://doi.org/10.1007/s42521-019-00014.
[13] Cui et al, 2019.
[14] Thompson, Andrew. All the News. *Kaggle* (2017). https://www.kaggle.com/snapcrack/all-the-news
[15] Ibid.
[16] https://ca.finance.yahoo.com/quote/%5EIXIC/history?p=^IXIC&.tsrc=fin-srch
[17] https://ca.finance.yahoo.com/quote/%5EGSPC/history?p=^GSPC&.tsrc=fin-srch
[18] https://ca.finance.yahoo.com/quote/%5ENYA/history?p=^NYA&.tsrc=fin-srch
[19] https://ca.finance.yahoo.com/quote/%5EDJI/history?p=^DJI&.tsrc=fin-srch
[20] https://ca.finance.yahoo.com/quote/%5EGSPTSE/history/

Publication	Open	Int	String
Author	High	Int	String
Date	Low	Int	Date
Year	Close	Int	Int
Month	Adjusted Close	Int	Int
Content	Volume	Int	String
Id	Int	Exchange Name	String
url	String		

Figure 1. Data Dictionary for Stock Market Data and News Headline Datasets

Approach



Obtain Data

To obtain the data two sources were used. The News Headlines datasets was obtained from the open source website Kaggle. Due to the size of the files, it was separated into three individual files. The second was sourced from Yahoo Finance, within the historical data search engine, the parameters were set to January 1st, 2015 to December 31st, 2017. The five markets were then searched, and the datasets individually downloaded.

Clean Data

The data sources were called in python to merge together into two dataframes, the first was called ‘News Headlines’ and the second ‘Stock Markets’. To clean the News Headlines data frame, the columns containing irrelevant features duplicates and tuples with missing dates were removed. To limit erasing articles with the same title but different content, duplicates where filtered by publication, headline and author. Data types were also corrected. The removal of duplicates and data with missing data lead to 17,442 articles being removed from the News Headlines data frame. This is an acceptable as it is less than 10% of the original dataset.

News Source Weighting

As there are many different News Outlets for each day sampled. The News outlets must be weighted in accordance to the theoretical number of people they reach. This will allow for us to determine how effective an article is. It will also assist in the training the predictive model by

assisting in the weighting of the overall sentiment scores, thus limiting over and under fitting. To accomplish this the Twitter followers for each news outlet will be taken for 2015, 2016 and 2017. The Friedman test will then be used to assign ranks to each outlet.

Sentiment Analysis

To accomplish the sentiment analysis the python library NLTK will be used. The tokenizer will then be used to remove whitespace. The remove noise function will be used to lemmatize, stem and clean the punctuation from the sentences. Stopwords will be removed and the tokenized words will then be put back into a dictionary for classification. A TF-IDF matrix will be made and then trained using the Naïve Bayes Classifier to assign a positive, negative and neutral sentiment to each article. Depending on model accuracy further preprocessing and feature engineering may be needed. If an accurate model cannot be produced lexicon-based analysis may be used.

Creating the Predictive Model

Once sentiment analysis is completed, the News Headlines and Stock Market Datasets will be merged and sorted based on date. This dataset will then be split into training and test sets. Pearson Correlation will be done to discover any feature correlations that need to be removed in order to streamline the model. Naïve Bayes, Random Forest and K Means clustering will be attempted to discover which model is the most accurate. Model accuracy will be determined by overall model accuracy, recall, precision and F score. Emphasis will be placed on minimizing recall value to allow for a more conservative model.

Visualization and Analysis

Upon completion of the model the data will be visualized using Seaborn and MatPlotLib packages in Python. Tableau will be used to make dashboards. A time series plot will be used to visual public sentiment and stock market index performance over time. A dashboard will also be made to dive deeper into how each individual publisher correlates to market performance. Ideally a correlation can be made to see how each market index is affected by the sentiment expressed.

References

Alanyali, Merve, Helen Susannah Moat, and Tobias Preis. “Quantifying the Relationship Between Financial News and the Stock Market.” *Scientific Reports* 3, no. 1 (2013). https://doi.org/10.1038/srep03578.

Audrino Francesco., Fabio Sigrist, Daniele Ballinari. The impact of sentiment and attention measures on stock market volatility, International Journal of Forecasting, 2019, ISSN 0169-2070,
https://doi.org/10.1016/j.ijforecast.2019.05.010.

Corsi, Fulvio and Roberto Renò (2012) Discrete-Time Volatility Forecasting With Persistent Leverage Effect and the Link With Continuous-Time Volatility Modeling, Journal of Business & Economic Statistics, 30:3, 368-380, DOI: 10.1080/07350015.2012.663261

Cui, Xin., Daniel Lam, and Arun Verma. “Embedded Value in Bloomberg News & Social Sentiment Data.” *Bloomberg Professional Services*, 2019. https://www.bloomberg.com/professional/sentiment-analysis-white-papers/.

Javers, Eamon. “Thomson Reuters Gives Elite Clients Early Edge.” CNBC. CNBC, June 12, 2013. https://www.cnbc.com/id/100809395.

Khedr, Ayman E., and Nagwa Yaseen. "Predicting stock market behavior using data mining technique and news sentiment analysis." *International Journal of Intelligent Systems and Applications* 9, no. 7 (2017): 22.

“Leading Indicators - Consumer Confidence Index (CCI) - OECD Data.” The OECD. Accessed February 24, 2020. https://data.oecd.org/leadind/consumer-confidence-index-cci.htm.

Luciana S. Malagrino, Norton T. Roman, Ana M. Monteiro,
Forecasting stock market index daily direction: A Bayesian Network approach, Expert Systems with Applications, Volume 105, 2018, Pages 11-22, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.03.039.

Ranco, Gabriele, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. “The Effects of Twitter Sentiment on Stock Price Returns.” *Plos One* 10, no. 9 (September 21, 2015). https://doi.org/10.1371/journal.pone.0138441.

Renault, T. Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digit Finance* (2019). https://doi.org/10.1007/s42521-019-00014.

Sprenger, T.O., Tumasjan, A., Sandner, P.G. and Welpe, I.M. (2014), Tweets and Trades: the Information Content of Stock Microblogs. Eur Financial Management, 20: 926-957. doi:10.1111/j.1468-036X.2013.12007.x

Williams, Christopher. “The $300m Cable That Will Save Traders Milliseconds.” The Telegraph. Telegraph Media Group, September 11, 2011. https://www.telegraph.co.uk/technology/news/8753784/The-300m-cable-that-will-save-traders-milliseconds.html.

Yin Ni, Zeyu Su, Weiran Wang, Yuhang Ying, A novel stock evaluation index based on public opinion analysis, Procedia Computer Science, Volume 147, 2019, Pages 581-587, ISSN 1877-0509.
https://doi.org/10.1016/j.procs.2019.01.212.

Yousuf, Hibah “Thomson Reuters Ends Privileged Access to Consumer Sentiment Data.” CNNMoney. Cable News Network. Accessed February 24, 2020. https://money.cnn.com/2013/07/08/investing/thomson-reuters-consumer-sentiment/index.html.