Analyzing Twitter Sentiment for Predicting Stock Market Movement

Utkarsh Kumar Singh
utksingh@iu.edu

Adarsh Bhandary
adnbhand@umail.iu.edu

School of Informatics and Computing
Indiana University Bloomington
IN, 47405, U.S.A.

Abstract

Stock Market prediction has been a well known topic of interest for quite a while. Now-a-days, the social media being able to represent the public sentiment and opinion about current events seems like a good topic of interest to gather data from. Especially, twitter has attracted a lot of attention from researchers for studying the public sentiments. Stock market prediction based on public sentiments expressed on twitter has been an intriguing field of research. The thesis of this work is to observe how well the changes in stock prices of a company, the rises and falls, are correlated with the public opinions being expressed in tweets about that company. Understanding author's opinion from a piece of text is the objective of sentiment analysis. The present paper have employed bag of words textual representations with extended features for analyzing the public sentiments in tweets. In this paper, we have applied sentiment analysis and supervised machine learning principles to the tweets extracted from twitter and analyze the correlation between stock market movements of 'Nike' and sentiments in tweets. In an elaborate way, positive news and tweets in social media about a company would definitely encourage people to invest in the stocks of that company and as a result the stock price of that company would increase. At the end of the paper, it is shown that a strong correlation exists between the rise and falls in stock prices with the public sentiments in tweets.

Keywords: Social Media Prediction, Sentiment Analysis, Twitter Sentiment Analysis, Supervised machine learning, Bag of words.

Introduction:

Earlier studies on stock market prediction are based on the historical stock prices. Later studies have debunked the approach of predicting stock market movements using historical prices. Stock market prices are largely fluctuating. The efficient market hypothesis (EMH) states that financial market movements depend on news, current events and product releases and all these factors will have a significant impact on a company's stock value [2]. Because of the lying unpredictability in news and current events, stock market prices follow a random walk pattern and cannot be predicted with more than 50% accuracy [1].

With the advent of social media, the information about public feelings has become abundant. Social media is transforming like a perfect platform to share public emotions about any topic and has a significant impact on overall public opinion. Twitter, a social media platform, has received a lot of attention from researchers in the recent times. Twitter is a micro-blogging application that allows users to follow and comment other users thoughts or share their opinions in real time [3]. More than million users post over 140 million tweets every day. This situation makes Twitter like a corpus with valuable data for researchers [4].Each tweet is of 140 characters long and speaks public opinion on a topic concisely. The information exploited from tweets are very useful for making predictions [5].

Rest of the paper is organized as follows. Section 2 describes the related works and Section 3 discusses the data portion demonstrating the data collection and pre-processing part. In Section 4 we discuss the sentiment analysis part in our work followed by Section 5 which examines the correlation part of extracted sentiment with stocks. In Section 6 we present the results, accuracy and precision of our sentiment analyzer followed by the accuracy of correlation analyzer. In Section 7 we present our conclusions and Section 8 deals with our future work plan.

Related Work

The most well-known publication in this area is by Bollen [10]. They investigated whether the collective mood states of public (Happy, calm, Anxiety) derived from twitter feeds are correlated to the value of the Dow Jones Industrial Index. They used a Fuzzy neural network for their prediction. Their results show that public mood states in twitter are strongly correlated with Dow Jones Industrial Index. Chen and Lazer [12] derived investment strategies by observing and classifying the twitter feeds. Bing et al. [15] studied the tweets and concluded the predictability of stock prices based on the type of industry like Finance, IT etc. Zhang [13] found out a high negative correlation between mood states like hope, fear and worry in tweets with the Dow Jones Average Index. Recently, Brian et al. [14] investigated the correlation of sentiments of public with stock increase and decreases using Pearson correlation coefficient for stocks. In this paper, we took a novel approach of predicting rise and fall in stock prices based on the sentiments extracted from twitter to find the correlation. The core contribution of our work is the development of a sentiment analyzer which works better than the one in Brian's work and a novel approach to find the correlation. Sentiment analyzer is used to classify the sentiments in tweets extracted. The human annotated dataset in our work is also exhaustive. We have shown that a strong correlation exists between twitter sentiments and the next day stock prices in the results section. We did so by considering the tweets and stock opening and closing prices of Nike over a year.

Data Collection and Cleaning

A. Data Collection

A total of 13198 tweets were collected from the period of one year from October 2016 to September 2017 on Nike was extracted using a tweet exporter python script (involving urllib and pyquery modules). The tweets were filtered using the words 'Nike' and 'Feel'. We have extracted tweets that represent the emotions of public about Nike and Nike products over a period of time. Stock opening and closing prices of Nike from October 2016 to September 2017 are obtained from Yahoo! Finance [23].

B. Data Pre-processing

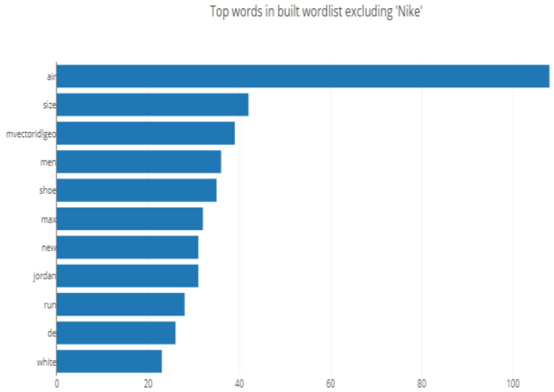
Stock prices data collected is not complete because of weekends and public holidays when the stock market is closed. The missing data is approximated using a simple technique by Goel [3]. Stock data usually follows a concave function. So, if the stock value on a day is x and the next value present is y with some missing in between. The first missing value is approximated to be (y+x)/2 and the same method is followed to fill all the gaps. Also, tweets consist of many acronyms, emoticons and unnecessary data like pictures and URL's. So, tweets are pre-processed to represent correct emotions of public. For pre-processing of tweets, we employed three stages of filtering: Tokenization, Stopwords removal and regex matching for removing special characters.

1) Tokenization: Tweets are split into individual words based on the space and irrelevant symbols like emoticons are removed. We form a list of individual words for each tweet.

2) Stopword Removal: Words that do not express any emotion are called Stopwords. After splitting a tweet, words like a, is, the, with etc. are removed from the list of words using stopwords corpus from NLTK module.

3) Regex Matching for special character Removal: Regex matching in Python is performed to match URLs and are replaced by the term URL. Often tweets consists of hashtags (#) and @ addressing other users. They are also replaced suitably. For example, #Nike is replaced with Nike and @CEO is replaced with USER. Prolonged word showing intense emotions like coooooooool! is replaced with cool! After these stages the tweets are ready for sentiment classification.

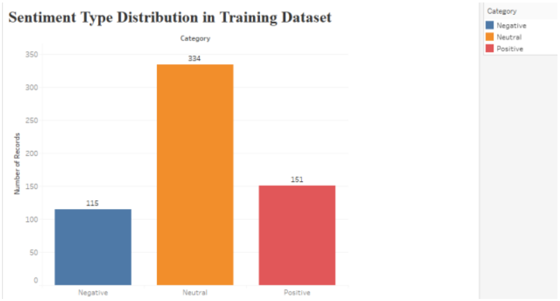
After that the wordlist was built by simple count of occurrences of every unique word after pre-processing across all the training dataset.



Sentiment Analysis and Methodology

Sentiment analysis task is very much field specific. There is lot of research on sentiment analysis of movie reviews and news articles and many sentiment analyzers are available as an open source. The main problem with these analyzers is that they are trained with a different corpus. For instance, Movie corpus and stock corpus are not equivalent. So, we developed our own sentiment analyzer.

Tweets are classified as positive, negative and neutral based on the sentiment present [18]. 600 tweets out of the total tweets are examined by humans and annotated as 'P' for Positive, 'NT' for Neutral and 'N' for Negative emotions. For classification of nonhuman annotated tweets a machine learning model is trained whose features are extracted from the human annotated tweets.



Above graph depicts sentiment type distribution in the training dataset.

A. Feature Extraction

Textual representations are done using the following method:

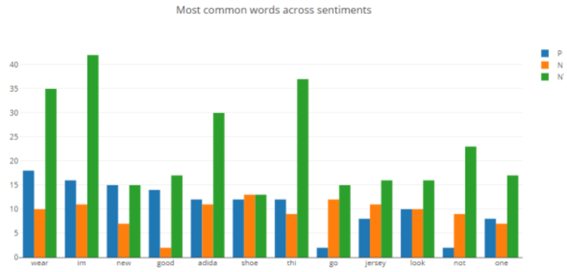
Bag of Words - The bag-of-words model is one of the simplest language models used in NLP. It makes an unigram model of the text by keeping track of the number of occurrences of each word. This can later be used as a features for Text Classifiers. In this bag-of-words model we only take individual words into account and give each word a specific subjectivity score. This subjectivity score can be looked up in a sentiment lexicon. If the total score is negative the text will be classified as negative and if its positive the text will be classified as positive. It is simple to make, but is less accurate because it does not take the word order or grammar into account.

In order to **not** push any other algorithm to the limit on the current data model, we added some extended features that help to classify tweets. A common sense suggests that special characters like exclamation marks and the casing might be important in the task of determining the sentiment. The following features was added to the data model:

Feature name	**Explanation**
Number of uppercase	people tend to express with either positive or negative emotions by using A LOT OF UPPERCASE WORDS
Number of !	exclamation marks are likely to increase the strength of opinion
Number of ?	might distinguish neutral tweets -

Feature name	**Explanation**
	seeking for information
Number of positive emoticons	positive emoji will most likely not occur in the negative tweets
Number of negative emoticons	inverse to the one above
Number of …	commonly used in commenting something
Number of quotations	commonly used in commenting something
Number of mentions	sometimes people put a lot of mentions on positive tweets, to share something good
Number of hashtags	just for the experiment
Number of urls	similar to the number of mentions

Extraction of these features was done before any pre-processing happened.

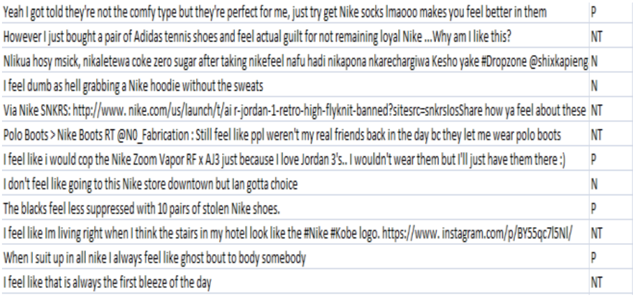


Above graph depicts the most common words across sentiments.

B. Model Training

All experiments were executed on a dual-socket, 12-core Intel Xeon system with 512GB of memory, available at Indiana University - Bloomington. As mentioned above, we used Python 3.6.1 with SciPy 0.18.1, NumPy 1.11.2, Pandas 0.19.0,and scikit-learn 0.18 for all coding in this work.

The features extracted using the above methods for the human annotated tweets are fed to the classifier and trained using supervised machine learning algorithms : Naive Bayes, Random Forrest & XGBoost Classifier. The results are almost comparable. Out of the three, model trained with Naive Bayes is picked because of its sustainability of meaning and promising performance over large datasets. The results of sentiment classification are discussed in the following sections. The devised classifier is used to predict the emotions of non-human annotated tweets.



Yeah I got told they're not the comfy type but they're perfect for me, just try get Nike socks lmaooo makes you feel better in them	P
However I just bought a pair of Adidas tennis shoes and feel actual guilt for not remaining loyal Nike ...Why am I like this?	NT
Nlikua hosy msick, nikaletewa coke zero sugar after taking nikefeel nafu hadi nikapona nkarechargiwa Kesho yake #Dropzone @shixkapieng	N
I feel dumb as hell grabbing a Nike hoodie without the sweats	N
Via Nike SNKRS: http://www. nike.com/us/launch/t/ai r-jordan-1-retro-high-flyknit-banned?sitesrc=snkrsIosShare how ya feel about these	NT
Polo Boots > Nike Boots RT @N0_Fabrication : Still feel like ppl weren't my real friends back in the day bc they let me wear polo boots	NT
I feel like i would cop the Nike Zoom Vapor RF x AJ3 just because I love Jordan 3's.. I wouldn't wear them but I'll just have them there :)	P
I don't feel like going to this Nike store downtown but Ian gotta choice	N
The blacks feel less suppressed with 10 pairs of stolen Nike shoes.	P
I feel like Im living right when I think the stairs in my hotel look like the #Nike #Kobe logo. https://www. instagram.com/p/BY55qc7l5Nl/	NT
When I suit up in all nike I always feel like ghost bout to body somebody	P
I feel like that is always the first bleeze of the day	NT

Table shows a sample of annotated tweets by the sentiment analyzer.

B.1. Naïve Bayes Classifier

A naive bayes classifier works by figuring out the probability of different attributes of the data being associated with a certain class. This is based on bayes' theorem. The theorem is

P(A|B) = (P(B|A),P(A)) / P(B)

This basically states "the probability of A given that B is true equals the probability of B given that A is true times the probability of A being true, divided by the probability of B being true."

We see that, Naive Bayes was our most accurate clasifier. We chose to use Bernoulli as our event model for the Naive Bayes classifier because our feature vector is using a bag of words model, where the values are binary. We considered using the Multinomial event model which counts word occurrences. However, tweets are such short texts that rarely contain multiple occurrences of the same word, thus Bernoulli was adequate for our project. A disadvantage would be that conditional independence often does not hold in reality for text, yet this model performed fairly well for our project.

The performance of this model is discussed in the coming sections.

B.2. Random Forrest Classifier

Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set which seemed a good approach for classifying sentiments in context to our project.

The performance of this model is discussed in the coming sections.

B.3. XGBoost Classifier

XGBoost (eXtreme Gradient Boosting) is an advanced implementation of gradient boosting algorithm. Standard Gradient Boosting has no regularization like XGBoost, therefore it also helps to reduce overfitting. It implements parallel processing and is blazingly faster as compared to GBM. It also allow users to define custom optimization objectives and evaluation criteria.

The performance of this model is discussed in the coming sections.

Correlation Analysis of Price and Sentiment

The stock price data of Nike are labeled suitably for training using a simple program. If the previous day stock price is more than the current day stock price, the current day is marked with a numeric value of 0, else marked with a numeric value of 1. Now, this correlation analysis turns out to be a classification problem. The total positive, negative and neutral emotions in tweets in a 3 day period are calculated successively which are used as features for the classifier model and the output is the labeled next day value of stock 0 or 1.The window size is experimented and best results are achieved when the sentiment values precede 3 days to the stock price. A total of 356 instances, each with 3 attributes are fed to the classifier with a split proportions of 90% train dataset and the remaining dataset for testing. The accuracy of the classifier is discussed in the results section.

Evaluation and Results

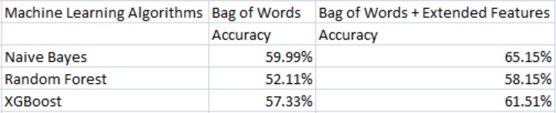
A. Baseline

As our lower bound, we set our baseline as randomly predicting 2 of the 5 sentiments, which gives 40% accuracy.

This following sections give an overview of accuracy rates of the trained classifiers.

B. Sentiment Analyzer Results

The above sections discussed the method followed to train the classifier used for sentiment analysis of tweets. The classifier trained on Random Forest algorithm with a split percentage of 70 for training the model and remaining for testing the model showed an accuracy of 58.15%. The XGBoost classifier model with same dataset showed an accuracy of 61.51%. The best results are obtained using the Naive bayes classifier which showed an accuracy of 65.15%. Therefore, the model trained with Naive Bayes algorithm is picked to classify the nonhuman annotated tweets because of its promising accuracy for large datasets. Numerous studies have been conducted on people and they concluded that the rate of human concordance, that is the degree of agreement among humans on the sentiment of a text, is between 65% and 75%. They have also synthesized that sentiment analyzers above 65% are very accurate in most of the cases. Provided this information, the results we obtained from the sentiment classification can be observed as good figures while predicting the sentiments in short texts, tweets, less than 140 characters in length.



Machine Learning Algorithms	Bag of Words	Bag of Words + Extended Features
	Accuracy	Accuracy
Naive Bayes	59.99%	65.15%
Random Forest	52.11%	58.15%
XGBoost	57.33%	61.51%

Table depicts the results of sentiment classification when trained with different machine learning algorithms.

C. Stock Price and Sentiment Correlation Results

A classifier is presented in the previous sections that is trained with aggregate sentiment values for 3-day period as features and the increase/decrease in stock price represented by 1/0 as the output. Total data is split into two parts, 90 percent to train the model and remaining for testing operations.

The classifier results show an accuracy value of 59.46% when trained using Logistic regression algorithm and the accuracy rate varied with the training set. When the model with SVM is trained with 90 percent of data, it gave a result of 62.16%. These results give a significant edge to the investors and they show good correlation between stock market movements and the sentiments of public expressed in twitter. This trend shows that with increasing dataset the models are performing well.

We would like to incorporate more data in our future work.

Conclusion

In this paper, we have shown that a strong correlation exists between rise/fall in stock prices of a company to the public opinions or emotions about that company expressed on twitter through tweets. The main contribution of our work is the development of a sentiment analyzer that can judge the type of sentiment present in the tweet. The tweets are classified into three categories: positive, negative and neutral. At the beginning, we claimed that positive emotions or sentiment of public in twitter about a company would reflect in its stock price. Our speculation is well supported by the results achieved and seems to have a promising future in research.

Discussion & Future Work

In this work, we have considered only twitter data for analyzing people's sentiment which may be biased because not all the people who trade in stocks share their opinions on twitter. Stocktwits [24] is a financial communication platform designed solely for sharing ideas and insights of investors, entrepreneurs and traders. The current study can be extended by incorporating Stocktwits data. In addition to this, data from news can also be included for an exhaustive public opinion collection.

While training the sentiment analyzer, 600 tweets are used which is comparatively a less number to train a sentiment analyzer. In future, we look forward to human annotate more than 5,000 tweets and train the classifiers. With increasing size of training datasets, the models tend to perform better.

We also look forward to experiment with other textual representations for feature extraction such as Word2vec and N-gram in future in order to improve the accuracy of our sentiment analyzer. All these remain as areas of future research.

Acknowledgement

We are extremely grateful to Professor Allen Riddell for his constant guidance and support.

References

[1] Qian, Bo, Rasheed, Khaled, Stock market prediction with multiple classifiers, Applied Intelligence 26 (February (1)) (2007) 2533, http://dx.doi.org/10.1007/s10489-006-0001-7.

[2] E.F. Fama, The behavior of stock-market prices, The Journal of Business 38 (1) (1965) 34105, http://dx.doi.org/10.2307/2350752

[3] J. Leskovec, L. Adamic and B. Huberman. The dynamics of viral marketing. In Proceedings of the 7th ACM Conference on Electronic Commerce. 2006

[4] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology. 2009.

[5] A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010, pp. 13201326

[6] Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: Proceedings of the ACM International Conference on Web Intelligence, pp. 492-499 (2010)

[7] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A.: Correlating financial time se-ries with micro-blogging activity. In: Proceedings of the fifth ACM international confer-ence on Web search and data mining, pp. 513-522 (2012)

[8] Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., Weber, I.: Web

search queries can predict stock market volumes. PLoS ONE 7(7), e40014 (2011)

[9] Gilbert, E., Karahalios, K.: Widespread Worry and the Stock Market. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp.58-65 (2010)

[10] Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Compu-tational Science, 2(1), 1-8 (2011)

[11] Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. ”Twitter catches the flu: detecting influenza epidemics using Twitter.” Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.

[12] R. Chen and M. Lazer, Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement, Cs 229, pp. 15, 2011.

[13] L. Zhang, Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation, pp. 130, 2013.

[14] Pagolu, V. S., Challa, K. N. R., Panda, G., & Majhi, B. (2016). Sentiment analysis of twitter data for predicting stock market movements. *arXiv preprint arXiv:1610.09225.*

[15] Bing, Li, Keith CC Chan, and Carol Ou. ”Public sentiment analysis in Twitter data for prediction of a company’s stock price movements.” e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on. IEEE, 2014.

[16] https://dev.twitter.com/overview/api

[17] Mittal, Anshul, and Arpit Goel. ”Stock prediction using twitter sentiment analysis.” Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf) (2012).

[18] Agarwal, Apoorv, et al. ”Sentiment analysis of twitter data.” Proceedings of the workshop on languages in social media. Association for Computational Linguistics, 2011.

[19] Tomas, M., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. Proceedings of Neural Information Processing Systems, Lake Tahoe, December 2013, 3111-3119

[20] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[21] Dickinson, Brian, and Wei Hu. ”Sentiment analysis of investor opinions on twitter.” Social Networking 4.03 (2015): 62.

[22]http://brnrd.me/social-sentiment-sentiment-analysis/

[23] https://finance.yahoo.com/quote/NKE/history?period1=1479099600&period2=1511758800&interval=1d&filter=history&frequency=1d

[24] http://stocktwits.com/home