Extract Stock Sentiment from Twitter data

Aditya Panchal
A. D. Patel Institute of Technology, India

Abstract— It is a well-known interest to predict stock market movements. Nowadays, social media is perfectly reprehensible for the public's mood and opinion on current events. In general, Twitter[1] has attracted a lot of attention from researchers to research the public's emotions. A fascinating field of research has been the stock market prediction based on public sentiments expressed on twitter. Previous studies concluded that the overall public mood collected from Twitter could well be correlated with the Dow Jones Industrial Average. The stock market prediction attempts to determine a company stock's future value. The company should make a profit if the future stock price can be predicted successfully. The aim of this research is to analyze how well a company's movements in stock prices, rising and falling, are associated with the public views expressed in the company's tweets. Knowing the perspective of the reader from a piece of text is the intention of analyzing sentiment. The present paper used textual representations to examine public feelings in tweets, Word2vec[2] for analysis of public sentiments in tweets. In this paper, they applied sentiment analysis and supervised machine learning concepts to tweets derived from twitter and explored the connection between a company's stock market movements and tweet feelings. Positive news and tweets about a company in social media will definitely encourage people to invest in that company's stocks.

Keywords: Sentiment Analysis, Natural Language Processing, Stock market prediction, Machine Learning, Word2vec, Python, Logistic Regression, Support Vector Machine, Random Forest

I. Introduction

Because of the underlying nature of the financial domain, stock prices are considered to be very dynamic and impressive to rapid changes. Financial researcher concludes that news articles, blogs, and stock market predictions are important topics in many business revenues. Previous stock market prediction experiments have based on historical share prices. The approach to forecasting stock market fluctuations using historical values has been refuted by later studies. Prices on the stock market fluctuate significantly. The Efficient Market Hypothesis (EMH) notes that financial market movements are dependent on media, current events and consumer leases, all of which will have a significant impact on the stock value of a company. An effective approach to the forecasting of values and share prices is the goal of stock market forecasters. Bourse prices fluctuate more than stock prices decrease and the stock price rises. Due to the volatile situation, bond prices adopt random patterns in current events and media, and can not be calculated more than 50% accurately. The goal is to produce great benefits by means of well-defined trade policies. The advent of social media has given rise to plenty of knowledge about pubic feelings. Social media is transformed as a perfect platform that expresses popular sentiments on every subject and has a huge impact on the general public. The general sentiment of the company's social data which can be important variables that influence the company's stock price. The various online social network platforms which make large amounts of data available. The analysis and reliability of a model can, therefore, improve the correlation of data from social media with the historical cost.

Recent researchers have been impressed with Twitter, a social media platform. Twitter is a microblogging application that allows users to follow and post on other users ' thoughts and share their opinions in real-time [3]. Every tweet has 140 characters and communicates to the public on a topic in brief. That is why Twitter is like a body with important researchers ' knowledge. The information used by tweets is very useful for forecasting. More than a million users post over 140 million [4] tweets every day. This situation makes Twitter like a corpus with valuable data for researchers [5].

In this paper, we contribute to the field of sentiment analysis of twitter data. Sentiment analysis is a study that addresses opinion-oriented natural language processing [7-second paper]. These opinion-based studies include emotion and mood recognition, rankings, calculations of relevance in the form of texts. The various analytical tools offered for calculating the feeling analysis of given data are G-POMS, N-gram, Lingmotif, LIWC, POMS, SentiStrength.

The problem description is to create a sentiment analysis model for inventory forecasting in a large-scale distributed setting. Apply clustering and SVM classification to the sentiment score to improve accuracy and apply the prototype to speed up output in a distributed environment. The dataset must be filtered with metadata such as a person's exact location, re-tweets, and the directories for the data set selected. Compare the calculation with the distributed environment Map-reduce. The dataset should be filtered with additional metadata such as an individual's exact location, the number of re-tweets, the number of followers in the chosen data set.

Several studies are available which include Twitter as a major public analysis source. Before its publication, Asur and Huberman [6] forecasted box office receipts for a movie on the basis of the public perceptions for Twitter movies. Together with Twitter Google flu patterns for predicting disease outbreaks are being widely studied. The twitter data to detect the flu outbreaks have been analyzed by Eiji[7] and Ruiz[8] used time-constrained charts to research the issue of correlating to flu outbreaks and Twitter microblogging with stock price shifts and volume trading respectively.

II. Previous work study

Bollen is the most common publication in this field. They examined the associated value of the Dow Jones Industrial Index in collective moods (Happy, Calm, Anxiety) from Twitter Feeds. For their prediction, they used a Fuzzy neural network. The results of these findings indicate the strong correlation of public moods with the Dow Jones Industrial Index on Twitter. In the analysis and classification of Twitter feeds, Chen and Lazer extracted investment strategies. The tweets were investigated by the Bing and the share prices based on industry category such as finance, IT etc concluded predictability.

For tweets with the Dow Jones Average Index, Zhang found a high negative relationship between mood states like optimism, fear, and concern. Brian has analyzed the connection between public feelings and increases for stock and decreases using the Pearson coefficient of inventory connection. In this paper, we have taken a new approach forecasting stock price rises and drops based on Twitter feelings in order to see the correlation. The main contribution of our work is the creation of an analyzer of sentiments that can predict rates more accurately. A sentiment analyzer is used to classify sentiments into extracted tweets. The human data set is also comprehensive in our work. We have seen that there is a strong correlation between Twitter feelings and the next day shares. We have done this by looking through one year at Microsoft's tweets and stock opening and closing prices

III. Dataset

Features like URL, mention, hashtag, recency and authors pieces of information plays a major role, hence those data's are considered during analysis for better outputs.

– URL: Most connected tweets typically provide a purposeful introduction to the links. Spam also includes links on Twitter. We also use a feature to indicate if a Tweet contains a connection in our template ranking.
– Mention: In a tweet, usually "@" is used to address other users prior to a username. It is possible that the text of the tweet is more of personal communication.
– Hashtag: A hashtag is a term that begins with a character "#" in the text of the Tweet. It will be used for the theme of the tweet.
– Recency: Twitter provides real-time text streams and is often supposed to be better for Tweet retrieval with more recent results.

For the social media platform, Twitter Further data from the author can also be used for spammer identification analysis.

1) Status: the author has always written about the involvement of an author in the number of tweets (status). Spammers who publish a large number of tweets are likely to be the most active writers. Thus, for the ranking of tweets, they use the number of statuses.
2) Place: The place associated with the tweet to classify the place to post the message. It includes country code, country code, Id, a place name.
3) Followers and friends: In Twitter, a user can choose to follow some other users for one reason or another which he or she finds interesting. When userA follows userB, the private stream of userA updates all Tweet posted by userB. They call userA a userB fan and userB a userA follower. The number of users indicates the user's popularity. The friend count also represents the user's type.
4) Retweets: Twitter offers user services for retweeting other users ' tweets. Each retweet starts with the @RT symbol. One feature in the spam detection system is the retweet of a user's newest tweets.
5) Listed: A user can divide his friends according to certain parameters into different lists. If a user is often listed, the tweets are very important for a large population of users. They use a feature that tests how often a tweet writer is included in the Tweet rating.

IV. Data pre-processing

The collected stock price data is not understandable because the stock market does not function on weekends or public holidays. Typically, inventory information follows a concave pattern. So, if the stock value on a day is x and you are without the next value. Approximately the first missing is (y+x)/2 and all holes are filled with the same form. Tweets contain various acronyms, pictures and URL's as well as needless data. Tweets are therefore ready to reflect the best social sentiments. We used three steps for sorting of preprocessing tweets: tokenization, removal of stopwords and regex matching for deletion of 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.
3) Regex Matching for special character Removal: Regex matching in Python is performed to match URLs and is replaced by the term URL. Often tweets consist of hashtags(#) and @ addressing other users. They are also replaced suitably. For example, #Microsoft is replaced with Microsoft and @Billgates 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.

V. Workflow (sentimental analysis)

The task of the feeling the study is very special to the field. Most scholars are available as an open-source to evaluate emotions in movie reviews and news articles. This analyzer has the major problem of having another corpus. For example, a film body and an inventory body aren't equal. We have also developed the analyzer of our own feelings. Tweets on the basis of the current feeling are rated as positive, negative and neutral. A plethora of tweets was analyzed by humans from the overall tweets and annotated as one positive tweet, 0 neutral tweet, and 2 critical tweets. A machine learning algorithm is equipped for the identification of annotated non-human tweets whose features are derived from human tweets.

A. *Feature Extraction:*

The representation Word2vec is much easier, more sophisticated and a new technique that operates by mapping words to a 300-dimensional matrix. Once each word is converted to a single vector, word vectors can be added up to create a corresponding vector for a given collection of words. In this form of representation, the relationship between the terms is maintained exactly. This sustained relationship between word concepts makes the word2vec model very attractive for textual analysis. Vectors of words differ between Rome and Italy are very close to the difference between vectors in France and Paris. The corresponding matrix, consisting of 300-dimensional vectors of all terms in a tweet, serves as features of that template in this representation.

B. Model Training:

Attributes extracted from the previous methods were fed into the cluster using random forest algorithms for the individual tweets annotated hereafter. The textual representations have both been successful and the results are comparable. Of those models trained in word2vec, the sustainable significance and promising performance of word2vec are selected over large datasets. In the following sections, the results of sentiment classification are explained. The designed classifier is used to predict the emotions of annotated non-human tweets.

C. Classifiers:

There are several classifiers that can be used and below there are two mentioned classifiers.

Support Vector Machine (SVM) [9] is mainly classifying by building hyperplane that divides cases belonging to various categories. A Vector Support Machine (SVM) is an algorithm for guided classification that has been successfully applied recently to text classification tasks.

$$c(x) = \begin{cases} 1 & w.\phi(x) + b \geq k \\ -1 & w.\phi(x) + b \leq -k \end{cases}$$

where, $w = \{w_1,...w_n\}$ is a weight vector.
$x = \{x_1,.....x_n\}$ is a input vector.
$\Phi(x)$ is kernel function.

Classification of the Random Forest [10] is tree-based. This consists of various category trees, which can be used to forecast the class mark on the basis of the categorically dependent variable of one data point. The error rate of this classifier relies on the comparison between any two forest trees, which contributes a certain or particular tree intensity in the forest. The trees should be high and the comparison should be as low as possible to minimize the error rate.

D. Regression Method:

Logistic regression [11] is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from the logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

VI. WORKFLOW(CORRELATION ANALYSIS OF PRICE AND SENTIMENT)

The Microsoft stock price data was suitably marked for a basic software practice. If the stock price for the previous day is higher than the current stock price, the present day shall be shown with a numerical value of 0, and the numerical value shall be 1. This analysis of correlations turns out to be a problem of classification. Production is the labeling of the next-day value of the stock0 and 1:1. The gap width is checked, with optimal results when the mood feelings precede 3 days of stock prices. The overall positive, negative and neutral emotions are measured successively in tweets over the 3-day period.

VII. RESULT

Each paragraph provides an overview of the reliability of classifiers qualified. All calculations are carried out on a virtual java machine by Weka tool [12].

A. Result of Sentimental Analysis:

The above sections addressed the approach used to train the classifier used to interpret the tweets. The classification with Word2vec features of human tweets, conditioned on the Random Forest Algorithm for the learning of the model and remaining for checking the model with a split percentage of 90, demonstrated the precision of 70,2 percent. Due to its promising reliability for large data sets and consistency in terms of word sense, the template trained for word2vec representations was chosen to identify nonhuman annotated tweets. Several experiments have been carried out and they have found that the level of agreement between humans about the feeling of a text is between 70% and 79%. In most cases, they have summed up the fact that sentiment analyzers over 70% are quite reliable. The results obtained from the emotions category can be found with this knowledge in the short messages, tweets with a length of fewer than 140 characters as very good figures and the estimation of sentiments.

B. Results of Correlation between Stock Price and Sentiment:

In the previous parts, a Classifier is provided with the inclusion of 3-day sentiment values as characteristics and an increase/decline in stock prices which is the product of 1/0. Maximum information is split into two parts, 80 percent for design learning and the remainder for research.

Once trained using the Logistic regression algorithm and the precision frequency varied with the training set the classification results show a precision value of almost 69 percent. When trained with 90% of the results, the LibSVM model provided a result of 71,82%. These results provide investors with a significant edge and reveal a good correlation between bond movements and the public sentiments expressed on Twitter. This phenomenon shows that the models do well with that data set. For our future work, we would like to include more information.

VIII. CONCLUSION

In this paper, we have shown that there is a strong correlation between the rise/fall in the company's stock prices and the public opinions or emotions expressed by tweets about that company. The main contribution of our work is the development of a sentiment analyzer using random forest that can judge the type of sentiment present in the tweet which is generally affecting the rise/fall of the stock price. Tweets are categorized into three categories: positive, negative and neutral. Initially, we suggested that positive feelings or public sentiment about a company on twitter would reflect positive stock price deviation. The results we achieve are strongly behind our speculation and appear to have a promising Research Future.

REFERENCES

[1] https://twitter.com/
[2] https://en.wikipedia.org/wiki/Word2vec
[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] 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)
[8] 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)
[9] https://en.wikipedia.org/wiki/Support-vector_machine
[10] https://en.wikipedia.org/wiki/Random_forest
[11] https://en.wikipedia.org/wiki/Logistic_regression
[12] https://www.cs.waikato.ac.nz/ml/weka/
[13] https://arxiv.org/pdf/1610.09225.pdf
[14] https://ijarcce.com/upload/2017/march-17/IJARCCE%20129.pdf