Stock Prediction Using Twitter Sentiment Analysis

Problem Statement

Stock exchange is a subject that is highly affected by economic, social, and political factors. There are several factors e.g. external factors or internal factors which can affect and move the stock market. Stock prices rise and fall every second due to variations in supply and demand. Various Data mining techniques are frequently involved to solve this problem. But technique using machine learning will give more accurate, precise and simple way to solve such issues related to stock and market prices.
"Stock Price Prediction Using Twitter Sentiment Analysis" a method for predicting stock prices is developed using news articles. 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. 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. A prediction model for finding and analysing correlation between contents of tweets and stock prices and then making predictions for future prices can be developed by using machine learning.

Background

Stock price prediction is one of the most important topic to be investigated in academic and financial researches. Various Data mining techniques are frequently involved in the studies. To solve this problem. But technique using machine learning/deep learning will give more accurate, precise and simple way to solve such issues related to stock and market prices.

On 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 user's thoughts or share their opinions in real time. More than million users post over 140 million tweets every day. This situation makes Twitter like a corpus with valuable data for researchers. 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. Sentiment analysis of twitter data and sentiment classification is the task of judging opinion in a piece of text as positive, negative or neutral.

In this project a method for predicting stock prices is developed using Twitter tweets about various company. Sentiment analysis of the collected tweets is used for prediction model for finding and analysing correlation between contents of news articles and stock prices and then making predictions for future prices will be developed by using machine learning.

Methodology

Step1: Data Collection

Tweets on Microsoft, Google, AAPL, are extracted from twitter API. The tweets will have collected using Twitter API and filtered using keywords like $ MSFT, # Microsoft,

#Windows etc. Not only the opinion of public about the company's stock but also the opinions about products and services offered by the company. The keywords used for filtering are devised with extensive care and tweets are extracted in such a way that they represent the exact emotions of public about Microsoft over a period of time. The news on twitter about Microsoft and tweets regarding the product releases can also be included. Stock opening and closing prices of Microsoft are obtained from Yahoo! Finance.

Step2: Data Pre-Processing

Stock prices data collected is not complete understandably because of weekends and public holidays when the stock market does not function. The missing data is approximated using a simple technique. 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.

Tweets consists 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, stop words 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 removed. Form a list of individual words for each tweet
2) Stop word Removal: Words that do not express any emotion are called Stop words. 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 are replaced by the term URL.

Step 3: Sentiment Analysis

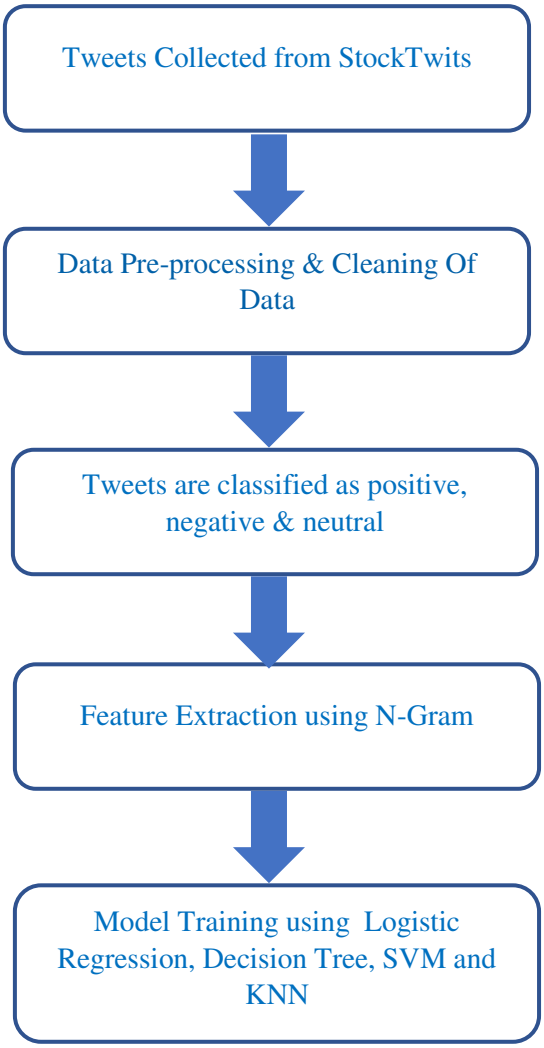
Sentiment analysis task is very much field specific. Tweets are classified as positive, negative and neutral based on the sentiment present. Out of the total tweets are examined by humans and annotated as 1 for Positive, 0 for Neutral and 2 for Negative emotions. For classification of nonhuman annotated tweets, a machine learning model is trained whose features are extracted from the human annotated tweets.

Step4: Feature Extraction

Textual representations can be done using n-grams. N-gram Representation: N-gram representation is known for its specificity to match the corpus of text being studied. In these techniques a full corpus of related text is parsed which are tweets in the present work, and every appearing word sequence of length n is extracted from the tweets to form a dictionary of words and phrases. For example, the text "Microsoft is launching a new product" has the following 3-gram word features: "Microsoft is launching", "is launching a", "launching a new" and "a new product". In our case, N-grams for all the tweets form the corpus. In this representation, tweet is split into N-grams and the features to the model are a string of 1s and 0s where 1 represents the presence of that N-gram of the tweet in the corpus and a 0 indicates the absence.

Step5: Model Training

The features extracted using the above methods for the tweets are fed to the classifier and trained using classification methods like Logistic Regression, Decision Tree, SVM and KNN to estimate the movement of the change in stock market price vs the volume as well as sentiment of news articles and tweets. Apply Linear Regression to find relation between the change in stock market price vs the volume as well as sentiment of news articles and tweets. Architecture is shown below:



Experimental Design

Dataset

1) Tweets from StockTwits
2) News articles from
    - IBM Alchemy Data News API
    - The Guardian API
    - NYTimes Article Search API
3) *Stock Information:*
    - *Google Finance API*
      Provides no delay, real time stock data in NYSE & NASDAQ
    - *Yahoo Finance API*

The updates are 15 minutes late but provides historical day-by-day stock data

Evaluation Measures

1 Measure correlation between
Volume of tweets vs change in stock price
Sentiment of tweets vs change in stock price
Volume of news articles vs change in stock price
Sentiment of news article vs change in stock price

2. Mean Squared Error for Linear Regression Model
Loss function and accuracy percentage for Classification model

Software and Hardware Requirements

Python based Computer Vision and Deep Learning libraries will be exploited for the development and experimentation of the project. Tools such as Anaconda Python, Jupyter Notebook and libraries such as OpenCV, Tensorflow, and Keras will be utilized for this process.

References

[1] S. A. R. Nai-Fu Chen and Richard Roll, Economic Forces and the Stock Market, The Journal of Business, vol. 59, no. 3, pp. 383–403, (1986).[Online]. Available: http://www.jstor.org/stable/2352710.

[2] E. F. Fama, Random Walks in Stock Market Prices, Financial Analysts Journal, vol. 51, no. 1, pp. 75–80, (1995).[Online]. Available: http://www.jstor.org/stable/4479810.

[3] S. J. Grossman and R. J. Shiller, The Determinants of the Variability of Stock Market Prices, National Bureau of Economic Research, Working Paper 564, October (1980) [Online]. Available: http://www.nber.org/papers/w0564.

[4] A. W. Lo and A. C. MacKinlay, Stock Market Prices do not Follow Random Walks: Evidence from a Simple Specification Test, Review of Financial Studies, vol. 1, no. 1, pp. 41–66, (1988).

[5] P. P¨a¨akk ¨onen and D. Pakkala, Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems,Big Data Research, vol. 2, no. 4, pp.166–186,(2015).[Online].Available: http://www.sciencedirect.com/science/article/pii/S2214579615000027.

[6] P. A. G. Xue Zhang and Hauke Fuehres, Predicting Stock Market Indicators through Twitter I Hope it is not as Bad as I Fear,Procedia – Social and behavioral Sciences, vol. 26, pp. 55–62, (2011).

[7] J. Bollen, H. Mao and X. Zeng, Twitter Mood Predicts the Stock Market, Journal of Computational Science, vol. 2, no. 1, pp. 1–8, (2011).

[Online]. Available: http://www.sciencedirect.com/science/article/pii/S187775031100007X.

[8] K. Mizumoto, H. Yanagimoto and M. Yoshioka, Sentiment Analysis of Stock Market News with Semi-Supervised Learning,In 2012 IEEE/ACIS 11th International Conference on Computer and Information Science (ICIS), pp. 325–328, May (2012).

[9] M. Z. F. Werner Antweiler, Is all that Talk Just Noise? the Information Content of Internet Stock Message Boards, The Journal of Finance,vol. 59, no. 3, pp. 1259–1294, (2004). [Online]. Available: http://www.jstor.org/stable/3694736.

[10] R. Ahuja, H. Rastogi, A. Choudhuri and B. Garg, Stock Market Forecast Using Sentiment Analysis, In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1008–1010, March (2015).

[11]"This Tweet Just Made Twitter's Stock Crash Hard | TIME." [Online]. Available: http://time.com/3839011/twitter-earnings-results/.

[12]"Forces That Move Stock Prices | Investopedia." [Online] Available:http://www.investopedia.com/articles/basics/04/100804.asp.

[13]"Support Vector Machines for Classification and Regression - SVM.pdf." [Online]. Available: http://trevinca.ei.uvigo.es/~cernadas/tc03/mc/SVM.pdf.

[14]S. Shen, H. Jiang, and T. Zhang, *Stock Market Forecasting Using Machine Learning Algorithms*.

[15]Nuno Oliveira, Paulo Cortez, and Nelson Areal. Progress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence, EPIA 2013, Angra do Hero´ısmo, Azores, Portugal, September 912, 2013. Proceedings, chapter On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume, pages 355–365. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013