International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Stock Value Estimation using Linear Regression

Rama Chandra Naradasu, Dharma Sri Harsha Tontepu, Subramanyam Kodukula

Abstract: ***Estimating the stock value is a tough task, because it depends more on value of stock and there`s no exact variable which is able to guess exactly a value of a stock every other day. What so Ever, Efficient Market Hypothesis said a stock vaue is crucially dependent upon new information. Many sources of info is the choice of the person in the social media.The choice of public on factory outlet from a specific firms can determine the stability of that particular firm and thus affect the decision of the many members to buy the company's stock. When using opinion as an important data, an appropriate analysis of that opinion is necessary. One of the well know example of using opinion as an important data is anbrief note of sentiments.Analysis of sentiment is a way to determine emotion within the choice of public about some reason, in the given case some of thecorporations goods. There is some way of analyzes of the sentiment required to guess stock prices. Bollen concludes on his research that with 87.6 per cent accuracy, interestin social networking site such as Twitter can guess DJIA interest. This shows a clear relationship between the analysis of sentiments and the stock values. Our goal in this research is to use simple sentiment analysis to forecast Indonesian stock market. Naïve Bayes and the algorithm Random Forest are used to identify tweet to measure a company's opinion. Sentiment analysis are used to display the stock value for the product. The prediction model is built using linear regression approach. Our research shows that predictive models are using before stock prices and hybrid feature as predictor provide the accurate prediction with determination coefficient of 0.9989 and 0.9983.***

Keywords : Sentiment Analysis ,Linear Regression, Stock Price, Supervised Learning, Efficient Market Hypothesis, Random Forest, Prediction.

I. INTRODUCTION

The most demanding task is predicting the stock prices. The reason is that there's no exact variable that can predict the stock price exactly each and every day. Based on an Efficient Market Hypothesis (EMH), recent information or recent stock prices is an important factor that effects more on a stock price changes. This knowledge, such as company news can change peoples mind and can effect people's option, whether to buy or neglect stock from the company. Many people are purchasing stock from the company so that the price is getting more. People are inclined to buy a reputable company. By seeing the connection between the firm and the customer one can know the stability of that firm.The seeing the usage of social networking sites forces a lot of business to build an official social media slot to keep in touch with their customers. This gives a chance to customers to easily give their opinion on the products. One of the social networking platforms that firms use is twitter. Many studies show social networking sites can influence the stock. Based on a study made by Johan Bollen, et.al, it concludes that some Twitter data mood states can predict the Dow Jones Industrial Average (DJIA) value with 87.6 per cent precision. A Study conducted by Anshul Mittal and Arpit Goel says if a value of DJIA can be predicted by twitter by 75.56 percent accuracy with the DJIA value, calmness and happiness mood conditions on previous days. This clearly shows that twitter information is very useful of determine the stock data.5th nation with the high number of active Twitter users is Indonesia,where accounts for 2.4 percent of all Twitter postings especially in jakarta. Since the above mentioned study was in acquaintance to English, the author was curious about the effects of Twitter data on Indonesian Firm stock value. This induces an interest to know more about the existing problem and to carry out this study. The application of the available classification and prediction algorithm to the dataset is a contribution of this research. The dataset consists of a report for the twitter and stock price. This twitter dataset is collected by many firms in Indonesia .

Manuscript received on April 02, 2020.
Revised Manuscript received on April 20, 2020.
Manuscript published on May 30, 2020.

* Correspondence Author
Rama Chandra Naradasu*, CSE department, Koneru Lakshmiah Educational Foundation, Vaddeswaram, India. ramanaradasu007@gmail.com
Dharma Sri Harsha Tontepu, CSE department, Koneru Lakshmiah Educational Foundation, Vaddeswaram, India, harshatontepu666@gmail.com
Dr. Subramanyam Kodukula, CSE department, Koneru Lakshmiah Educational Foundation, Vaddeswaram, India, smkodukula@kluniversity.in

II. DATASET

This study utilized two data types. In Indonesia, data is listed by stock values of multiple firms and the data contains choices on certain output produced by the firms. Through Twitter the choices were shared. The Firms with high-variating stock value and are already popular outlets in Indonesia are selected. The firms investigating were as follows:

Table 1.

Companies List
Exxon Mobil
Lukoil
Royal Dutch Shell
Total
Chevron
Reliance Oil Corp
China National Petroleum Corp

Data pooling was done in two weeks, from February 4th 2020 to February 19th 2020.

A.Google Graph Analysis

From Twitter Search API Google graph analysis (GGA) was derived. The company's official Twitter account and the company's product names were the keywords used for query parameters. Retrieved tweets were at Bahasa. Data recovered was formatted to JSON. To calculate the purpose of the sentiment analysis we gather tweet ids, time the tweet was posted and user who posted the tweet, and tweet andstatus etc.

B.Public Domain Dataset Analysis

Using Pubic Domains(Government Websites) CSV APIstock price dataset was gathered. Each day the Info being gathered were the open stock value and close stock value of the firms.

The data gathered was in CSV format. To create prediction modelthis data is being utilized as the predicted value and also coupled with the result of sentiment analysis.

III. LEXICAL ANALYSIS

The process for classifying the polarity of opinion is an Analysis of sentiment. Lexicon of sentiment is one more important thing for determining polarity. Sentiment lexicon is a word or sentence which contains a sentence's emotion .Lexicon for feelings are classified to two kinds. There are both positive and negative lexicons. Positive feeling or emotion is expressed byPositive lexicon(e.g. excellent, marvellous, extravagant, etc.) and negative feeling or emotion is expressed by Negative Lexicon (e.g. worst, imbecile, incorrigble, etc.). From the data gathered in this researchLexicon sentiment has been retrieved.First the collected data were tokenized into a monotonus word. Formal word was chosen which was categorized as lexiconusing the Indonesian dictionary definition. Whereas the Unethical words are tested manually by looking at each word . Later they are separated by the +ve lexicon and the -ve lexicon.

There are various kinds of classifications when determining polarity of choice . Choice is classified into 6 charecter states (Collected, Warning, Certain, Vital, Kind and Happy), opinion is classified into 3 groups (+ve, -ve, neutral) etc. One among many methodologies that can be utilized in sentiment analysis is Classifying by supervised learning. The method classifies opinion based on data / features that can be derived. Forextracting and using in classification there are many other features besides lexicon. Many of them include:

- Word and it`s impact or weight
- Parts of Speech (POS)Tagging
- Sentiment Shifter

Many algorithms may be used in supervised classification. Naïve Bayes,Support Vector Machine (SVM), Random Forest and Neural Network (with single-layer perceptron), Decision Tree are all used in this research algorithm.

A.Support Vector Machine (SVM)

SVM is an instance-based methodology, creating a linear function that optimizes class distance[7]. The algorithm uses instance data at the edge of class to build the class function. Data from this instance is called support vector. The SVM illustration below is.

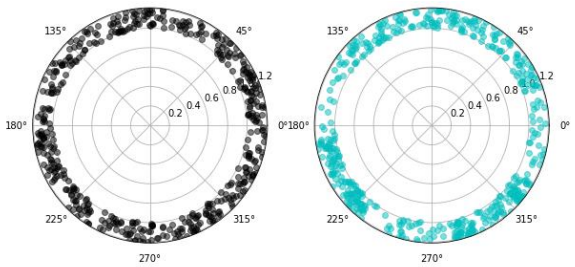

Fig 1.Illustrating Dataset comparison between Google Graph Analysis and Government Sources

B. Naïve Bayes

Naïve Bayes is a statistical classification methodology. It assume all instance data features are independent. To determine which class for every instance data belongs the algorithm uses the dependent probability. Below is Naïve Bayes ' probabilistic function.

$$P(C|F_1,F_2,F_3,\dots F_n) = \frac{P(F_1|C)P(F_2|C)P(F_3|C)\dots P(F_n|C)P(C)}{P(F_1,F_2,F_3,\dots F_n)} \quad (1)$$

- $P(C|F_1,F_2,F_3, \dots F_n)$ is probability of one instance classified as C given feature combination $F_1,F_2,F_3, \dots F_n$
- $P(F_1|C)$ is probability of feature F_1 given data classified as C
- $P(C)$ is probability of data classified as C
- $P(F_1,F_2,F_3, \dots F_n)$ is probability of feature combination $F_1,F_2,F_3, \dots F_n$

The instance data is being classified into class with higher probability value.

C.Decision Tree

Decision Tree is an algorithm based on a tree that represents a shape of a tree. Three parts of a tree exist. Core, inner and leaf node[9].[9]. Instance data are used as the root and the internal node, while the leaf node is the class. That root or inner node split represents the value of the split node function. Following is the decision tree diagram.

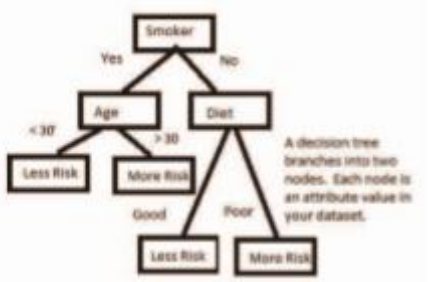

Fig. 2. Illustrating a decision tree (Source: http://www.refactorthis.net/)

It is a complication to identify which function is used for increasing depth as root and internal node. Selection of high-info function is also one of the features. The data benefit is a value ranging from 0 to 1, it indicates that the data classification function is very. If you get more information about 1 it indicates that the function is very important. The gain of information can be determined using the following equation

$$InfoGain(C,f) = Entropy(C) - Entropy(C|f) \quad (2)$$

- C is class
- f is feature which information gain value is going to be checked

Below is function to calculate class entropy value.

$$Entropy(C) = -\sum_{j=1}^{n} prob(j)\log_2 prob(j) \quad (3)$$

- C is class
- n is number of classes
- j is values of class

Below is function to calculate entropy of one feature

$$Entropy(C|f) = \sum_{j=1}^{n} prob(f=j)Entropy(K|f=j) \quad (4)$$

- C is class
- n is number of value in the feature
- j is feature values

D.Random Forest

Random Forest is a classification technique consisting of many model classifications based on grouping data into trees. Every tree can be formed by fissioning the number of characteristics for each division without pruning[8].

The characteristics are selected randomly by substitution.Every single data will be categorized by all trees after a random tree number has been established. As the main classifier is the tree that correctly classifies the majority of instance data. That's what vote is called.

IV. PREDICTION MODEL

In this research three things were predicted. They are prediction of price alterations, predictive percentage margins and estimation of price. Prediction of price alteration employed by controlled system of learning categorization.

Margin percentage and estimation of prices using linear regression, As the expected value can not be categorized. Linear regression is one among regression methods that is used for categorizing numerical instances [1]. The linear function is created by calculating weight values (w) for each feature (β).

$$x = \beta_0 w_0 + \beta_1 w_1 + \beta_2 w_2 + \cdots + \beta_n w_n$$

For a single instance data, X is the regression value. Here is an illustration of linear regression to have clear understanding of linear regression

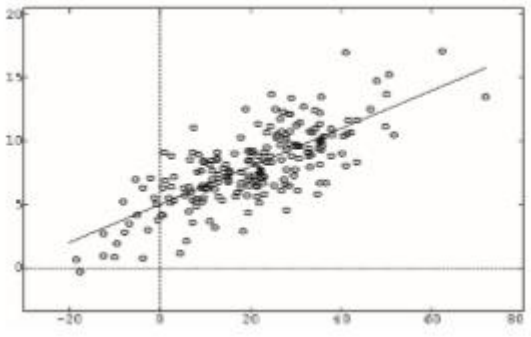

Fig. 3. Exemplification of Linear Regression

As shown on top exemplification, there is a linear line representing data distribution. The difference to the linear line of data from each instance is called residual. The linear function is created by searching for the appropriate strength value for each feature to reduce error (mean error) of every instance data.

Many ways of assessing a model of linear regression. Model that fits most data has a normal residual distribution within it. In addition, it may also be evaluated by checking the determination value coefficient (R2). R2 is a square of the correlation coefficient (R) between the values expected and the real values. R2 varied between 0 and 1.

V. EXPERIMENTAL DESIGN

A.Analytical Design

Purpose is to produe models that are able to categorize the tweet data's polarity. The approach used was by supervised learning classification. The cluster of Tweets were divided into three groups in this research (positive, negative, and neutral). Features used in classification were sentiment lexicon and sentiment shifters

The classification algorithms used in Chapter III described some. Two model classifications were further divided into three grades, with the highest precision. The percentile in positive tweet for every day was determined once all tweets were marked. The formula for the calculation was as follows:

$$\% = \frac{\#positive\_tweet}{\#positive\_tweet + \#negative\_tweet}$$

B.Predictive Model Design

a.Value Fluctuation Predction

The goal is to determine whether a firm's inventory price will increase or decrease on the observation day compared with the previous day's stock price. The indicator is the dominant percentage of tweets on the day before (ranging from 1 - 5 days). The method of supervised classification was followed by classification algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT). The model was tested to forecast stock value fluctuation by the tweet data which has already been categorized according to two sentiment classification models.

b.Margin Percentile Prediction

It is aimed at forecasting a company's share profit percentage on the observation day, using previous day knowledge as basis. To create the prediction model a linear model is used. The algorithm was linear (multinomial) regression. For this prediction, there were three linear models. Each of the features used was different. The linear model of all three models is shown Models vary from 1 - 5 days prior to day of observation. Tests with two types of tweet data are listed as previously explained in two ways for models with a good tweet output. On the basis of determination coefficient value the following models are evaluated.

1. $M_x = w_0 + \sum w_i M_{x-i}$
2. $M_x = w_0 + \sum w_i T_{x-i}$
3. $M_x = w_0 + \sum w_i M_{x-i} T_{x-i}$

- M_x : margin percentage on day x
- T_x : positive tweet percentage on day x
- w_0 : intercept weight
- w_i : feature weight

VI. EXPERIMENTALL RESULT

A. Result of Analysis

The above analysis classification is done for four methodologies mentioned before to create a model.The accuracy of this model is given below

Table 2.

Algorithm	Accuracy
Support Vector Machine (SVM)	38.64%
Naïve Bayes	**56.50%**
Decision Tree	46.02%
Random Forest	**60.39%**

B.Result of Prediction Model

The precision of the model for the mentioned algorithms are given below

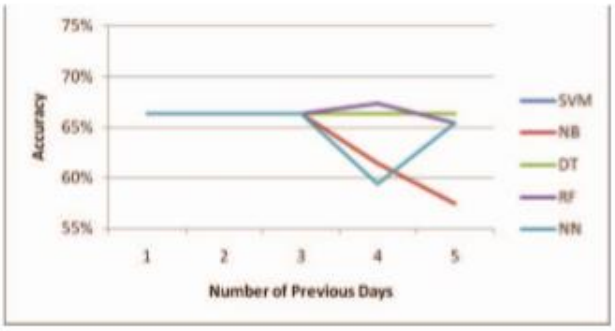

Fig. 4. Accuracy of Value Fluctuation Prediction

C.Margin Percentile Prediction Model

The R*R estimation of this model is described in the below graph

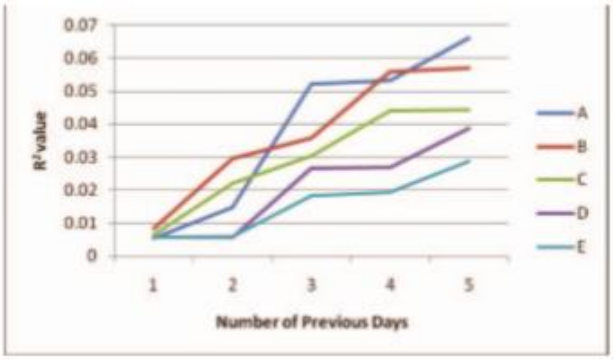

Fig. 6. Plot of R*R value of Margin Percentile Prediction Model

D.Price Prediction Model

The R*R value of the model for each feature useare described below

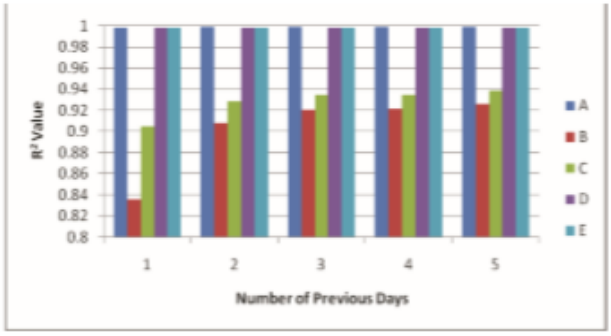

Fig. 7. Plot of R2 value of Stock Price Prediction Model

The output shows that the highest R2 value was created on five previous days using a prediction model with a price characteristic, which is 0.9989. The prediction model is followed by a 0.9983 hybrid feature five days earlier. All the models generated produced nearly 1 R2 values.

VII. CONCLUSION AND FUTURE WORK

Several points can be inferred from the findings of this study

- The created Random Forest Analyzes Model is able to classify Tweet Data with accuracy 60.39 percent, with tweet Data 56.50 percent with Naïve Bayes algorithm.
- The prediction of price fluctuation helps existing models to predict when the next price will rise or fall at 67.37% for Naïve Bayes ' classified tweet data and 66.34% for the Random Forestry classified tweet data.
- The model produced has a R^2 value near 0 in the margin percentage prediction. This implies that very few data were generated by produced models.
- The value of R^2 created models is close to 1 in pricing prediction. It means that many data fitted into the produced models. The highest R*R value is obtained from the previous model price.
- The R*R value of the model reduces the percentage of positive tweets

The suggestions for possible future works

- Improving the model for sentiment analysis by using another classification function.
- Extending the span of data collection within one month or more, such as POS tagging, word weighting, etc.
- Using a different approach to construct non-linear prediction models

REFERENCES

1. Investopedia. (n.d.). Efficient Market Hypothesis: Is The Stock Market Efficient? Retrieved June 24, 2015, from Investopedia:http://www.investopedia.com/articles/basics/04/022004.asp
2. Bollen, J., Mao, H., & Zeng, X. J. (2010). Twitter mood predicts the stock market. arXiv .
3. Stock Price Prediction Through the Sentimental Analysis of News Articles .978-1-7281-1340-1/19/$31.00 ©2019 IEEE
4. Mittal, A., & Goel, A. (2009). Stock Prediction Using Twitter Sentiment Analysis. CiteSeerX .
5. Berita 8. (2013, November 21). Ini 5 Negara Pengguna Aktif Twitter Terbanyak di Dunia. Retrieved June 25, 2015, from Berita 8: http://www.berita8.com/berita/2013/11/ini-5-negara-pengguna-aktiftwitter-terbanyak-di-dunia
6. Stock Price Prediction using Linear Regression based on Sentiment Analysis.978-1-5090-0363-1/15/$31.00 c 2015 IEEE
7. IOSR Journal of Business and Management (IOSR-JBM) e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 19, Issue 9. Ver. VI. (September. 2017), PP 24-33 www.iosrjournals.org
8. Stock Market Analysis: A Review and Taxonomy of Prediction Techniques.Int. J. Financial Stud. 2019, 7, 26; doi:10.3390/ijfs7020026 www.mdpi.com/journal/ijfs

AUTHORS PROFILE

Rama Chandra Naradasu, CSE department, Koneru Lakshmiah Educational Foundation, Vaddeswaram, India. ramanaradasu007@gmail.com.
An Aspirant who is wishing to reach the possible heights in the career.To be so trying to different fieds and research areas where he can find his abilities useful

Dharma Sri Harsha Tontepu, CSE department, Koneru Lakshmiah Educational Foundation, Vaddeswaram, India, harshatontepu666@gmail.com.
An Aspirant who is wishing to reach the possible heights in the career.To be so trying to different fieds and research areas where he can find his abilities useful.

Dr.Subramanyam Kodukula, CSE department, Koneru Lakshmiah Educational Foundation, Vaddeswaram, India, smkodukula@kluniversity.in. An Inspiring Faculty who works for the welfare of the students and tries to help the student any possible doubts and problems. He is a phenomenal and an inspirational guide.He gives us an ample amount of support and work that`ll guide us to the goal we try to achieve