Stock Price Forecasting using the Machine Learning Based on the Historical Stock Prices

Dr. Vimalkumar B. Vaghela[1], Tushar A. Champaneria[2], Hitesh D. Rajput[3]

[1, 3]Assistant Professor, Department of Computer Engineering, L. D. College of Engineering, Ahmedabad, Gujarat, India
[1]vimalvaghela@gmail.com, [3]hitesh.rajput@ldce.ac.in
[2]Assistant Professor, Department of Computer Engineering, Government Engineering College, Modasa, Gujarat, India
[2]tushar.1985@gmail.com
Corresponding Author: - Dr. Vimalkumar B. Vaghela

Abstracts: Forecasting stock price is an imperative financial subject matter that has involved researchers' attention for numerous years. The successful stock price forecasting helps to profit the company or individual otherwise it makes the problem of loss. Stock forecasting involves an assumption that basic information publicly accessible in the past has some predictive relationships to the future stock returns. This proposed research work tries to help the company and investors in the stock market to come to a decision for better timing for buying or selling stocks based on the information extracted from the historical prices of such stocks.

Keywords: Machine Learning, Data Transformation, Data Classification, Decision Tree, Future stock return, data mining techniques, Bombay Stock Exchange

1. INTRODUCTION

The macroeconomic atmosphere and the financial market are evolutionary, complex, non-linear, non-parametric dynamical systems. Two most important aspects about the stock are the: 1) Stock exchange and 2) Stock market. Stock exchange or sometimes people uses the name security exchange is the facility where traders and stockbrokers can buy and sell securities. Like, as shares of stock and bonds and other financial instruments. Stock exchanges also make possible for the issue and redemption of securities and other financial appliances including the imbursement of income and dividends [1].

Stock forecasting is very crucial key factors for investors in the stock market. The precise forecasting of stock is very demanding task and helps to investors for making the profit. Data Mining based approaches like predictive methods play very crucial role to analyze and forecast the stock price. Prediction is one of the important aspects of data mining and using in different areas especially in financial domain.

Stock market is affected by many highly inter-related economic, political, and sentimental aspects, which often interrelate with one another in a very complex way. As such, it has been always very much difficult to predict the movements of the stock price and stock market index.

The stock price depends on its inherent value and expected return on investment [2]. For that it requires to analyze the company's operations and the market in which the company is operating. It is probable to find out the company's inherent value and expected returns. Accordingly, the stock price can be predicted reasonably well. The methods that had been used to forecast stock market prices mainly characterize into three categories: fundamental analysis, technical analysis and traditional time series forecasting.

Fundamental analysis is a good approach only for long term basis. Conversely, for the short term and medium term assumptions, fundamental analysis is generally not suitable. It is also complicated to formalize fundamental analysis for automated decision support because the interpretation of financial analysis is often extremely subjective.

Technical analysis refers to the various approaches that aim to forecast future price movements using past stock prices and volume information. It is based on the hypothesis that history repeats itself and that future market directions can be determined by tentative historical price data. As such, it is supposed that price trends and patterns subsist that can be identified and exploited for profit. Most of the methods used in technical analysis are exceedingly subjective in nature and have been shown not to be statistically valid. Many of them even not have a rational explanation for their use [3].

Traditional time series predicting techniques in statistics have also been applied to forecasting stock price movements. Regression is one of the widely used techniques for forecasting the stock price movements; it is possible to model historical price values data as a non-linear function using recurrence relation. Based on the derived relation of the recurrence, it is used to forecast the future price. In general, time series forecasting is enhanced appropriate to short term forecasting, in general less than a year but it relies on a huge amount of high quality data.

This paper focuses on the technical analysis approach. Stock market prices data is present the pattern, so it is possible to use data mining approaches to discover these patterns in an automated manner. Once these patterns have been determined, future prices of the stock can be forecast.

Most challenging task of using data is to produce useful rules from raw data in a dataset/database for users to make decision, and these rules may be hidden extremely in the raw data. The problem with forecasting stock market price is that the amount of data is too large and huge. This paper focuses one of the data mining methods; which are the classification method on the past data available to try to find the investors to build their decision on whether to buy or sell that stock in order to attain financial profit.

The paper focuses is to analyze the past data available of the stock using the decision tree, Naïve Byes, Bagging, Random forest approaches as are the classification techniques used in data mining to forecast whether company or investor to know when to buy or sell their stock.

2. Related Work

Lin (2004) [4] proposed a approach to modify the filter rule by integrating three decision variables associated with fundamental analysis. An experimental test, using the stock of electronics companies in Taiwan, showed her method outperforms the filter rule. Lin considers in work the criteria for clustering the trading points which involved only the past information.

Upadhyay A. et al (2012) [5] uses the multi logistic regression model for determining the aspects which considerably influence the performance of the company in the stock market. Binary logistic regression is used to make the relationship among the stock performance and financial ratios.

Patil et al (2015) [6] uses the support vector machine for the stock market prediction. SVM algorithm works on the large dataset value which assembled from various global financial markets. Correlation is the important factor to understand the relationship between the market stock index and global markets.

Recently, data mining techniques and artificial intelligence methods like rough set method, artificial neural network, decision tree have been applied to this area. Data mining refers to extracting or mining knowledge from large data sets [7].

Deshmukh, B.G., et al (2016) [8], have presented a system that can forecast the stock market movements which are based on the past stock prices and market sentiment analysis. They used data of standard and poor's 500 (S&P 500) from Yahoo finance. They have used Naïve Bayes classification for sentiment analysis and stock movement were predicted using support vector machine, logistic and neural network methods.

Bini, B.S. et al (2016) [9] have used clustering and multiple regression for forecasting the stock price. Clustering is performed on stock data obtained from NSE, which makes the name of the finest companies as output. Then comparison between partitioning based, hierarchical, model based and density based methods are carry out with the help of validation index such as c-index, Jaccard index, rand index and silhouette index [10].

Zhang and Zhou (2004) [11], they have discussed KDD in the reference to financial application. They referred that developing data mining techniques have shown huge potentials in financial applications and will continue to flourish in the new knowledge-based economy. They list the data mining algorithm are Association rule mining (ARM) uncovers attractive connection patterns among a large set of data items by showing attribute-value conditions that occur together frequently. Classification and prediction is models that explain and differentiate data classes or concepts, Clustering analysis segments a huge set of data into subsets or clusters. Each cluster is a collection of data objects that are similar to one another within the same cluster but dissimilar to objects in other clusters. Sequential pattern and time-series mining looks for patterns where one event (or value) leads to another later event (or value).

Kumar and Valli (2013) [12], they introduce RFPID-algorithm to mine regular frequent Patterns using vertical data format which achieves better with large number of transactions and long Item sets with one database scan. The other advantage of vertical data format is, it uses simple operations like Unions, intersections, deletions, simple arrays etc., and also it judges non frequent and non regular item sets before generating the candidate sets.

Prasanna and Ezhilmaran (2013) [13] In this paper authors have discussed numerous attempts made by Researches for stock price prediction. These works show that data mining techniques can be applied for evaluation of past stock prices and acquire valuable information by estimating suitable financial indicators. There are two types of prediction methods are implemented by several researches to generate useful extracts. They are fundamental approaches and technical indicator based approaches. This tock data is highly volatile and unpredictable which makes data mining and AI techniques as suitable once. They suggest that Back propagation algorithm for training and suitable AI technique applied on some fundamental approaches may render promising results.

3. Methodology

Stock market is basically non-linear in nature. Forecasting of stock market price plays crucial role in stock business. Different data mining, neural network, fuzzy logic approaches are to be used to forecast the stock market price. Data mining can be applied on past and present data to generate the useful patterns and decision making system.

3.1 Methods for stock price prediction:

Fundamental Analysis:

It's mainly focuses with company rather than the stock. The analyst takes their opinion based on past performance of company like earning forecast etc.

Technical Analysis:

This method concerns with the purpose of stock price based on the past data values patterns of the stock. Different approaches can be used to generate interesting patterns using support vector machine, time series analysis, average methods. Obviously, the result of the technical analysis method is much better than the fundamental analysis method.

At the beginning, the data collected contained 9 attributes; this number was reduced manually to 6 attributes as the other attributes were found not important and not having a direct effect on the study. Table1 shows the 6 attributes selected with their descriptions and their possible values. The class attribute is the investor action whether to buy or sell that stock and it is named, "Action".

Table 1: Attribute Description

Attribute	**Description**	**Possible Values**
Previous	Previous day close price of the stock	Positive, Negative, Equal
Open	Current day open price of the stock	Positive, Negative, Equal
Min	Current day minimum price of the stock	Positive, Negative, Equal
Max	Current day maximum price of the stock	Positive, Negative, Equal
Last	Current day close price of the stock	Positive, Negative, Equal
Action	The action taken by the investor on this stock	Buy, Sell

At the starting, when the data was collected, all the data values of the attributes selected were continues numeric values. Data selection and data transformation was applied by generalization data to a higher level concepts so all the values became discrete. Previous day closing price of the stock was the criteria for transforming the each numeric attribute to discrete attribute values. If the values of the attributes open, max, min, last were greater than the value of attribute previous for the same trading day, the numeric values of the attributes were replaced by the value Positive. If the value of the attribute open, max, min, last were less than the value of the attributes previous then the numeric values of the attributes were replaced by Negative. If the values of those attributes were equal to the value of the attribute previous, the values were replaced by the equal. Table 2 shows a sample of the continues numeric values of the data before selecting the six attributes manually and before generalizing them to discrete values, while table 3 describes the same sample after selecting the six attributes and after transforming them to discrete values.

Table 2: Sample of stock price data before data selection and data transformation

Previous	**Open**	**Max**	**Min**	**Last**	**Action**
312	316	318	311	313	Sell
313	313	314	307	312	Sell
312	312	317	311	316	Buy
316	316	318	313	317	Buy
317	318	338	317	333	Sell
333	336	341	331	334	Sell
334	335	343	328	341	Sell
341	342	342	335	340	Sell
340	339	344	335	343	Buy
343	344	345	338	341	Sell

Table 3: Sample of stock price data after data transformation and data generalization

Previous	**Open**	**Max**	**Min**	**Last**	**Action**

Positive	Positive	Positive	Negative	Positive	Sell
Positive	Positive	Positive	Negative	Negative	Sell
Negative	Equal	Positive	Negative	Positive	Buy
Positive	Equal	Positive	Negative	Positive	Buy
Positive	Positive	Positive	Positive	Positive	Sell
Positive	Positive	Positive	Negative	Positive	Sell
Positive	Positive	Positive	Negative	Positive	Sell
Positive	Positive	Positive	Negative	Negative	Sell
Negative	Negative	Positive	Negative	Positive	Buy
Positive	Positive	Positive	Negative	Negative	Sell

3.2 Navie Bayes Classifier

Naive Bayes classifier is a classification algorithm based on Bayes' Theorem.A Naive Bayesian method is simple to construct, with no complex iterative parameter inference which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does unexpectedly well and is extensively used because it often outperforms more sophisticated classification methods.

Algorithm:

Bayes theorem provides a way of calculating the posterior probability, *P*(*c*|*x*), from *P*(*c*), *P*(*x*), and *P*(*x*|*c*). Naive Bayes classifier assumes that the effect of the value of a predictor (*x*) on a given class (*c*) is independent of the values of other predictors. This assumption is called class conditional independence.

Likelihood | Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability | Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where,

- *P*(*c*|*x*) is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- *P*(*c*) is the prior probability of *class*.
- *P*(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.
- *P*(*x*) is the prior probability of *predictor*.

Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

3.3 Bagging

Bagging is one of the ensemble classifier in data mining. An ensemble classifier is a technique which uses or combines several classifiers to enhance robustness as well as to achieve an improved classification performance from any of the constituent classifiers. Furthermore, this technique is more elastic to noise compared to the use of a single classifier. This method uses a 'divide and conquer approach' where a compound problem is decomposed into multiple sub-problems that are easier to understand and solve. Ensemble approaches have the advantage that they can be made to adapt to any changes in the monitored data stream more precisely than single model techniques. An ensemble classifier has better accuracy than single classification techniques in most of the cases. The success of the ensemble approach depends on the diversity in the individual classifiers with respect to misclassified instances. Bagging, which means bootstrap aggregation, is one of the simplest but most successful ensemble methods for improving unbalanced classification problems. For example, weak classifiers, such as decision tree algorithms, can be unstable, especially when the position of a training point changes slightly and can lead to a very different tree. This method is usually applied to decision tree algorithms, but it also can be used with other classification algorithms such as naïve bayes, nearest neighbor, rule induction, etc. The bagging technique is very useful for large and high-dimensional data, such as intrusion data sets, where finding a good model or classifier that can work in one step is impossible because of the complexity and scale of the problem.

3.4 Decision Tree

Decision trees can be used for various machine learning applications. But trees that are grown really deep to learn highly irregular patterns tend to over-fit the training sets. Noise in the data may cause the tree to grow in a completely unexpected manner. Random Forests overcome this problem by training multiple decision trees on different subspaces of the feature space at the cost of slightly increased bias. This means that none of the trees in the forest sees the entire training data. The data is recursively split into partitions. At a particular node, the split is done by asking a question on an attribute. The choice for the splitting criterion is based on some impurity measures such as Gini impurity or Shannon Entropy.

Gini impurity is used as the function to measure the quality of split in each node. Gini impurity at node N is given by:

$$G(N) = 1 - (P_1)^2 - (P_{-1})^2 \quad (1)$$

where P_i is the proportion of the population with class label i. Another function which can be used to judge the quality of a split is Shannon Entropy. It measures the disorder in the information content. In Decision trees, Shannon entropy is used to measure the unpredictability in the information contained in a particular node of a tree (in this context, it measures how mixed the population in a node is). The entropy in a node N can be calculated as follows:

$$S(N) = -P_1 \log(P_1) - P_{-1} log(P_{-1}) \quad (2)$$

where d is number of classes considered and P(w_i) is the proportion of the population labeled as i. Entropy is the highest when all the classes are contained in equal proportion in the node. It is the lowest when there is only one class present in a node (when the node is pure).

The obvious heuristic approach to choose the best splitting decision at a node is the one that reduces the impurity as much as possible. In order words, the best split is characterized by the highest gain in information or the highest reduction in impurity. The information gain due to a split can be calculated as follows:

$$Gain = I(N) - plI(N_L) - prI(N_R) \quad (3)$$

where I(N) is the impurity measure (Gini or Shannon Entropy) of node N, I(NL) is the impurity in the left child of node N after the split and similarly, I(NR) is is the impurity in the right child of node N after the split; NL and NR are the left and right children of N respectively' pL and pR are the proportions of the samples in the left and right children nodes with respect to the parent node.

Note that Equations 1, 2 and 3 can be used when there are only two splits at each node. However, depending on the algorithm used, there may be more than two splits and hence these formulae may be generalized as:

$$G(N) = 1 - \sum_{j=1}^{n} P_j$$

$$S(N) = -\sum_{j=1}^{n} P_j \log(P_j)$$

$$Gain = I(N) - \sum_{j=1}^{n} p_j I(N_j) \quad (4)$$

Algorithm 1: Decision Tree

input: X $(x_i, y_i)_1^n$ is the labeled training data
l is the current level of the tree
M is the set of features used to grow a tree

output: A tree which is configured to predict the class label of a test sample

```
l ← l + 1
CL← null;
CR← null;
/* CL ,CR are the left and right children of this node respectively */

MaxGain← 0;
/* MaxGain stores the value of the maximum possible gain that can be achieved from splitting a node– based on this, the final split is determined */

for j in M do
Sort((x_i, y_i)_1^n in the increasing order of the j^th feature
for i = 1 to n do
   CL← X[0:i];
   CR←X[i+1:n];
   Gain←GiniGain((x_i, y_i)_1^n[j], CL ,CR);
   if Gain >MaxGain then
     MaxGain← Gain
     CL ,CR   ←cL ,cR;
end
end
if CL does not satisfy the desired level of purity then
     DecisionTree(CL, l, M);
if CR does not satisfy the desired level of purity then
DecisionTree(CR, l, M);

/* A node is not pure when it has samples belonging to more than one class. In such a case, it needs to be split in turn. In this way, through successive function calls, the feature space is recursively partitioned. */
```

Algorithm 2: GiniGain

```
input:        X (x_i, y_i)_1^n is the labeled training data
     CL is the left child node
     CR is the right child node

output:       The Gini gain of the current split

GN←  Gini impurity of root node;
GL←  Gini impurity of left child;
GR←  Gini impurity of right child;
Ln←   number of samples in cL;
Rn←   number of samples in cR;
PL←   Ln=n ;
PR←   Rn=n ;
return GN - (PL X GL) - (PR X GR);
```

3.5 Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. Working of Random Forest Algorithm We can understand theworking of Random Forest algorithm with the help of following steps −

- Step 1 − First, start with the selection of random samples from a given dataset.
- Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- Step 3 − In this step, voting will be performed for every predicted result.
- Step 4 − At last, select the most voted prediction result as the final prediction result.

The following diagram will illustrate its working −

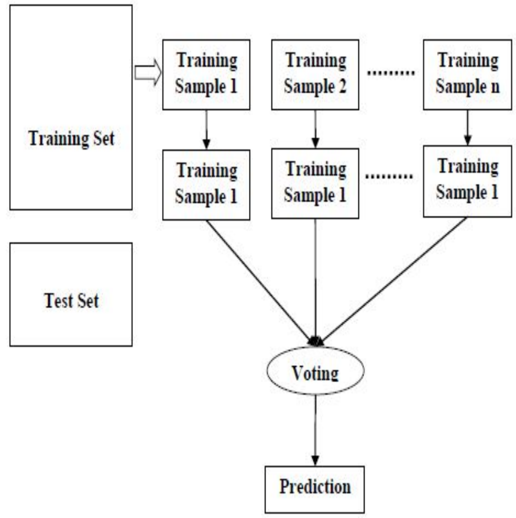

Figure 1. working of Random forest classifier

4. RESULT AND DISCUSSION

Once the data has been prepared and transformed, the next step was to build the classification model using the decision tree induction method. The decision tree method was selected because the building of decision tree classifiers does not require any domain knowledge, thus it is suitable for exploratory knowledge finding. Decision tree method is also suitable to handle the high dimensional data. The information gain ration is used to assign the rank to attributes and then it used to build the tree were every attribute is situated according to its information gain ratio.

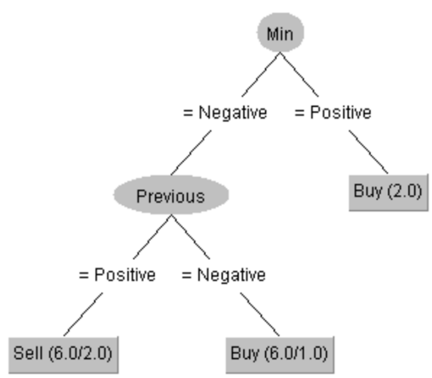

Figure 2. Sample decision tree for the YesBank

Table 5: classification accuracy using different classifiers

Classifier	Total Instance	Accuracy
Naïve Bayes	423	72%
Bagging	423	78%
Decision Tree	423	76%
Random Forest	423	82%

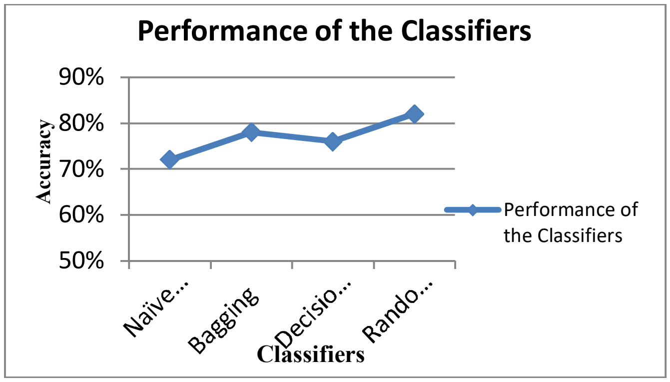

Figure 3. Performance of the various classifiers on YESBank dataset

Table 5 describes the accuracy of the classifier on the YESBank data. As we can see that Random forest gives better result compare to the other classifiers. In this paper, we use the 10-fold cross validation methods for the sample selection. There are other factors also play important role for predicting the stock price are such as internal financial factor like; financial reports, the overall performance of the market and news about the company. Also the external factor also affect such as: political decisions, political events, etc.

5. CONCLUSIONANDFUTUREWORK

This paper presents a proposal to about various approaches for predicting the stock price of any stock in the market. In this paper, uses different well-known classifiers such as decision tree, Naive Bayes, Bagging, and Random Forest as classifiers on the past data values of the stock price to predict the stock price which gives indication to the investor about whether he or she buy or sell in the stock market. Such proposed model can be helpful for company or investors to take the right decision regarding their stocks in order to derive and predictive information from the past data. The results for the proposed model were not ideal because so many other factors such as political events, investors' expectations influence and general economic conditions not covered. As for the future work, there is a scope of room for testing and enhancing the proposed model by evaluating the model over the whole companies listed in the stock market.

REFERENCES

[1] Stringham, Edward Peter (2015). Private Governance: Creating Order in Economic and Social Life. Oxford University Press.ISBN 9780199365166.
[2] Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K., and Wong, T.L. (2007) "Design and implementation of NN5 for Hong Kong stock price forecasting", Engineering Applications of Artificial Intelligence, 20, pp. 453-461.
[3] Ritchie, J.C., (1996) Fundamental Analysis: a Backtothe-Basics Investment Guide to Selecting Quality Stocks. Irwin Professional Publishing.
[4] Lin, C. H. "Profitability of a filter trading rule on the Taiwan stock exchange market." Masters, Department of Industrial Engineering and Management, National Chaio Tung University (2004).
[5] Upadhyay, Arun, Gautam Bandyopadhyay, and Avijan Dutta. "Forecasting stock performance in indian market using multinomial logistic regression." Journal of Business Studies Quarterly 3.3 (2012): 16.
[6] Patil, Sachin Sampat, Kailash Patidar, and Megha Jain. "Stock market prediction using support vector machine." International Journal of Current Trends in Engineering & Technology 2.1 (2016): 18-25.
[7] Han, J., Kamber, M., Jian P. (2011). "Data Mining Concepts and Techniques". San Francisco, CA: Morgan Kaufmann Publishers.
[8] Deshmukh, B.G., Jain, P.S., Patwardhan, M.S. and Kulkarni, V., 2016, August. Spin-offs in Indian stock market owing to twitter sentiments, commodity prices and analyst recommendations. In Proceedings of the International Conference on Advances in Information Communication Technology & Computing (p. 77). ACM.
[9] Bini, B.S.; Mathew, T. Clustering and Regression Techniques for Stock Prediction. Procedia Technol. 2016, 24, 1248–1255.
[10] Han, J., Kamber, M., Jian P. (2011). "Data Mining Concepts and Techniques". San Francisco, CA:Morgan Kaufmann Publishers.
[11] Vaghela, Vimalkumar B., Bhumika M. Jadav, and M. E. Scholar. "Analysis of various sentiment classification techniques." International journal of Computer applications 140.3 (2016): 975-8887.
[12] Zhang, Dongsong, and Lina Zhou. "Discovering golden nuggets: data mining in financial application." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 34.4 (2004): 513-522.
[13]Kumar, Vijay, and V. Valli Kumari. "Incremental mining for regular frequent patterns in vertical format." *Int. J. Eng. Tech* 5.2 (2013): 1506-1511.
[14]Vaghela, Vimal B., Amit Ganatra, and Amit Thakkar. "Boost a weak learner to a strong learner using ensemble system approach." *2009 ieee* international advance computing conference. IEEE, 2009.
[15]Prasanna, S., and D. Ezhilmaran. "An analysis on stock market prediction using data mining techniques." *International Journal of Computer* Science & Engineering Technology (IJCSET) 4.3 (2013): 49-51.
[16] Gandhi, Riya A., and Vimalkumar B. Vaghela. "Computing textual semantic similarity for short texts using different similarity measures." *Computing textual semantic similarity for short texts using different similarity measures* (2017).

First Author Dr. Vimalkumar B. Vaghela, is an Assistant Professor in Department of Computer Engineering at L. D. College of Engineering, Ahmedabad, Gujarat, India. He had done his Ph.D. in Relational Data Mining. He has published more than 37 research papers in international journals and conferences. He had published a book titled "Ensemble Classifier in Data Mining". He has more than 18 years of teaching experience; his biography was Biography has been published in the Who's Who in Science and Engineering – 11thEdition, 2010-11, 2013-14, 2016-17. Written a Book with title "Operating System" in Dreamtech Publisher. Senior Professional Members of Association for Computing Machinery (ACM), Life time Professional Member of Computer Society of India (CSI) & Professional Member of International Association of Engineers (IAENG).

Second Author Tushar A. Champaneria, is an Assistant Professor in Department of Computer Engineering at Government Engineering College, Modasa. He is pursing Ph.D. in area of IoT and He has published more than 32 research papers in international journals and conferences. He had published two patents. He has 3 years of industry experience and more than 13 years of teaching experience.

Third Author Hitesh D. Rajput, is an Assistant Professor in Computer Engineering Department at L. D. College of Engineering, Ahmedabad. He has more than 18 years of teaching experience.

DOI: https://doi.org/10.15379/ijmst.v10i1.3670