A Summer Internship Project Report

On

"Financial News Impact on Indian Stock Markets"

Carried out at the

Institute for Development and Research in Banking Technology, Hyderabad

Established by 'Reserve Bank of India'

Submitted by

Rishabh Miglani

Roll No. – 13MA20054

Integrated MSc. (Mathematics and Computing)

Indian Institute of Technology, Kharagpur.

Under the Guidance of

Prof. V. Ravi
Professor & Head,
Centre of Excellence in Analytics,
IDRBT, Hyderabad

May - July, 2016

CERTIFICATE

This is to certify that the summer internship project report entitled **"Financial News Impact on Indian Stock Markets"** submitted to **Institute for Development & Research in Banking Technology [IDRBT], Hyderabad** is a bonafide record of work done by **"Rishabh Miglani", Roll no. - 13MA20054, Integrated MSc. Mathematics and Computing, 2013-18, Indian Institute of Technology, Kharagpur"** under my supervision **from "6th May, 2016" to "9th July, 2016"**

(Project Guide)
Prof. V. Ravi
Professor & Head,
Centre of Excellence in Analytics,
IDRBT, Hyderabad

INDEX:

Content Page No.

1. **Abstract……** **1.**
2. **Introduction…….** **2.**
3. **Text Mining…….** **5.**
    a. **Data Collection.**
    b. **Data Representation and Preprocessing.**
    c. **Data Imputation.**
    d. **Analysis and Feature Selection.**
    e. **Model Creation using Machine Learning techniques.**
    f. **Result Analysis.**
4. **Proposed Methodology……** **7.**
    a. **Data Acquisition and Dataset Description.**
    b. **LIWC.**
    c. **Data Imputation.**
    d. **Feature Selection Techniques.**
    e. **Machine Learning techniques**
5. **Results……** **12.**
    a. **Results with all extracted features.**
    b. **Results with Chi-squared feature selection technique.**
    c. **Results with MRMR feature selection technique.**
6. **Conclusions……** **15.**

1.Abstract

Financial time series analysis has emerged as one of the important topics of interest over time. The basic definition of financial time series analysis explains the process of asset valuation over time. It has been evident from the past research that published news articles have effects on stock prices and investment decisions, indicating these have predictive power to affect stock markets. Most of the advances done in this field involves distributed lag models i.e., regression model for time series data used to predict current values of a dependent variable based on both the current values of an explanatory variable and the lagged values of this explanatory variable. The proposed model in this paper uses only the textual data in the form of published news articles to prepare a model efficient enough to predict stock prices in Indian financial markets. This approach involves the following steps: **1.** Extraction of news articles and historical stock prices corresponding to companies in Indian stock market. **2.** Processing of extracted news documents using LIWC (**L**inguistic **I**nquiry and **W**ord **C**ount) tool. **3.** Preparation of a model solely on the basis of these semantic textual features. Different models like like SVR, QRRF, GMDH, GRNN, RPART, RF and MLP were tested to obtain relevant results in the prediction analysis. Among all methods Neural Networks models are giving best results, especially GMDH and GRNN performing better than other models.

2.Introduction

Stock market prediction is the process of estimating future value of a company stock or other financial instrument traded on an exchange. “The efficient-market hypothesis suggests that stock prices reflect all currently available information about that stock” – Prof. Eugene Phama, Economist.

The information publicly available for stock market analysis includes historical stock price data, published news articles, expert recommendations and reaction of general folks on social media. Most of the research in this field focuses on using quantitative data while ignoring the enormous amount of qualitative textual data flooding the electronic and print media. But,

“The key to making a good forecast is not in limiting yourself to quantitative information”

– 2012 - Nate Silver, Statistician and Writer.

The efficient-market hypothesis posits that stock prices are a function of information and rational expectations, and that newly revealed information about a company's prospects is almost immediately reflected in the current stock price. This would imply that all publicly known information about a company, which obviously includes the information about the events related to an organization and other available information like historic prices, would already be reflected in the current price of the stock. Accordingly, changes in the stock price reflect release of new information, changes in the market generally, or random movements around the value that reflects the existing information set. Burton Malkiel, in his influential 1973 work ***“A Random Walk Down Wall Street”***, claimed that stock prices could therefore not be accurately predicted by looking at price history. As a result, Malkiel argued, stock prices are best described by a statistical process called a "random walk" meaning each day's deviations from the central value are random and unpredictable.

Despite the early criticism in this approach, recent advances in the field of financial forecasting have presented us with results much better than a random guess. Prediction methodologies fall into three broad categories which often overlap. They are fundamental analysis, technical analysis (charting) and technological methods.

1. **Fundamental Analysis:** It is based on evaluation of a company's past performance as well as the credibility of its accounts. In other words, Fundamental analysis is built on

the belief that human society needs capital to make progress and if a company operates well, it should be rewarded with additional capital and result in a surge in stock price.

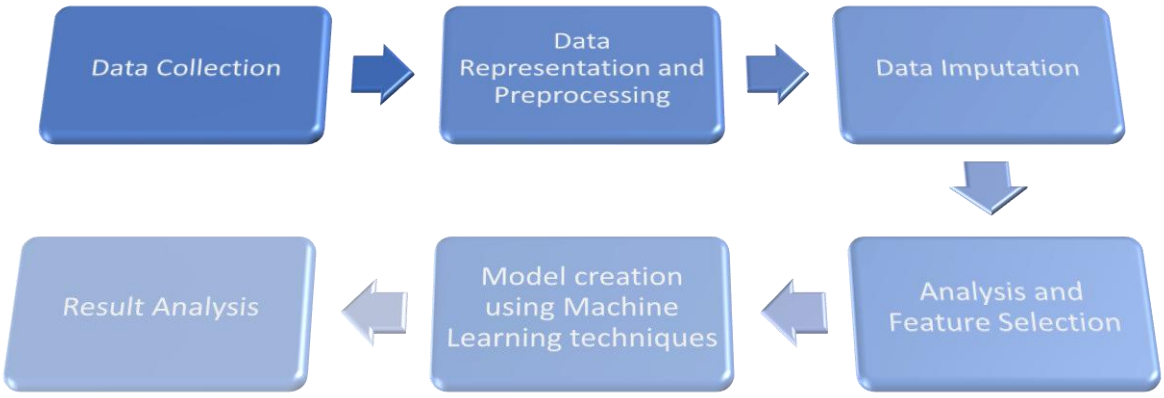
2. **Technical Analysis:** Future price of a stock is determined solely on the basis of (potential) trends of the past price (a form of time series analysis) employing patterns such as the head & shoulders or cup & saucer as well as statistical techniques like exponential moving average (EMA).
3. Technological Methods: Since the last decade, stock market prediction has moved into the technological realm. The most prominent techniques involve the use of artificial neural networks (ANNs) and Genetic Algorithms. The most common form of ANN in use for stock market prediction is the feed forward network utilizing the Back propagation algorithm with gradient descent to update respective weights in different layers of the network. Another form of ANN that is more appropriate for stock prediction is the time recurrent neural network (RNN) or time delay neural network (TDNN). Nowadays, the majority of academic research groups studying ANNs for stock forecasting seem to be using an ensemble of independent ANNs methods more frequently, with greater success.

Due to the rapidly increasing amount of textual data available and the range of interesting and important problems arising in text analysis, there has been growing interest in applying machine learning methods to text. Using textual data for financial forecasting proves to be a good option for increasing prediction accuracy. **Schumaker and Chen** employ a predictive machine learning approach for financial news article analysis that uses several different textual representations. They find that their model containing both article terms and stock price performs best in stock price prediction. They also find that among the textual representations, proper nouns outperform noun phrases which in turn outperform the de facto standard, the bag of words (BOW). **Lavrenko et al.** devise a language model that can characterize some financial trends. Associating news stories with forthcoming trends, the model learns that words such as loss, shortfall, and bankruptcy are most likely to precede a downward trend in the stock price, whereas merger, acquisition, and alliance are likely to be followed by an upward trend. This model produces noticeably better predictions for all trend types than the popular vector-space model. The models mentioned above utilized the syntactic textual features along with historic stock prices for model preparation. This approach has a drawback in the sense that features selected for the model preparation cannot

be standardized and moreover some of the syntactic features may not show any relevance in the future news articles. One of the alternatives is to use semantic textual features along with syntactical measures which can standardize the number and type of features to be used for model preparation corresponding to each stock. Tools for analyzing text on the basis of sentiments, emotions, opinions and moods are available which can score a document on a set of fixed features which can be used for model implementation. **Bollen et. al. (2010)** analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy) along with historic price data to predict changes in DJIA closing values. The proposed research in this project uses another similar tool called LIWC in order to assess given text on the basis of different emotions, thinking styles, social concerns, and even parts of speech. In addition to this, our proposed methodology is solely based on these textual features without any use of available quantitative information to predict stock prices. The results obtained after model implementation with different machine learning techniques were compared before and after the application of feature selection techniques in order to improve prediction accuracy.

3.Text Mining

The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information is extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. In the most general terms, text mining is a process of "turning text into numbers" (meaningful indices), which can then be incorporated in other analyses such as predictive data mining projects, the application of unsupervised learning methods (clustering), etc. The general text mining methodology can be explained in the form of the following process.



1. **Data Collection:** Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes.
2. **Data Representation and Preprocessing:** Real world textual data is generally incomplete, noisy and inconsistent. Therefore, it requires pre-processing. Tasks in data pre-processing includes Data cleaning (fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies), Data integration (using multiple databases, data cubes, or files), Data transformation (normalization and aggregation), Data reduction (reducing the volume but producing the same or similar analytical results) and Data discretization (part of data reduction, replacing numerical attributes with nominal ones). Data representation refers to the process of converting textual features to

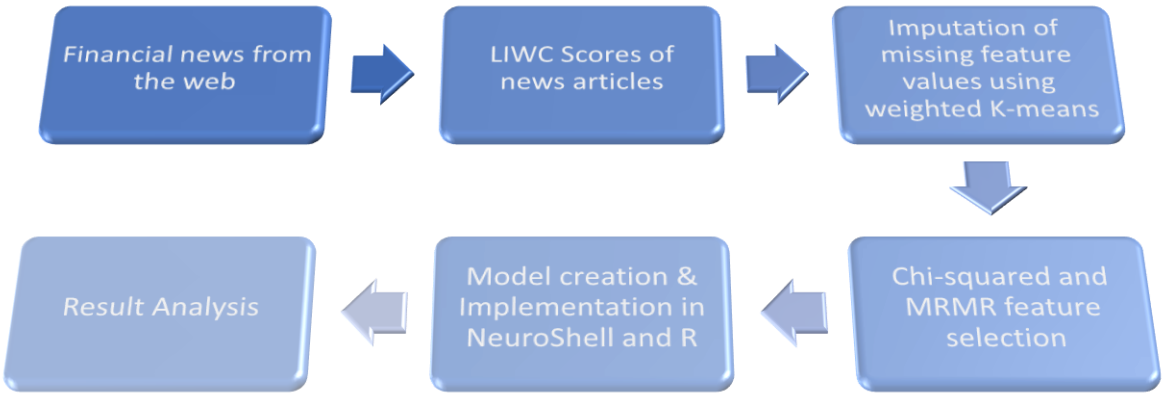
numeric features to make the information contained in the text accessible to the various data mining algorithms.

3. **Data Imputation:** In statistics, imputation is the process of replacing missing data with substituted values. The major "missingness mechanisms" include Missingness completely at random, Missingness at random, Missingness that depends on unobserved predictors, Missingness that depends on the missing value itself. There are two ways in which missing data is handled: 1. Deleting the incomplete records 2. Use of statistical or machine learning techniques to replace missing values with related values depending upon the values in complete records.
4. **Analysis and Feature Selection:** In machine learning and statistics, feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for three reasons: simplification of models to make them easier to interpret by researchers/users, shorter training times, enhanced generalization by reducing over fitting (formally, reduction of variance).
5. **Model creation using Machine Learning techniques:** This process involves application of machine learning algorithms to create model for fitting the acquired and processed textual data. Majority of the machine learning tasks can be divided into supervised, unsupervised and reinforcement learning tasks.

6. **Result Analysis:** This is the process of analysing the results obtained after application of above steps. This is done through various statistical measures including accuracy, mean squared error etc.

4.Proposed Methodology

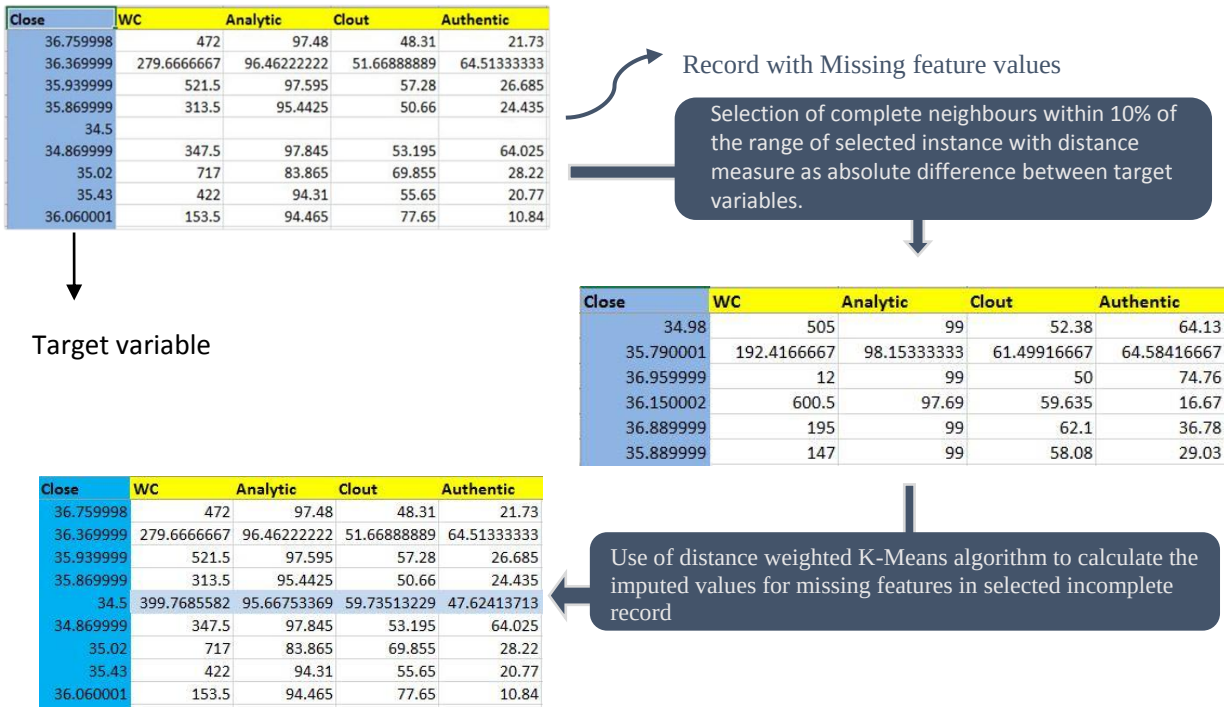
The figure given below explains in brief the model architecture prepared for the proposed research.



1. **Data Acquisition & Dataset Description:** Data collection for the proposed research was done through ***"Business Standard"*** online news resource **(Source: http://www.business-standard.com)**. Data was collected for 7 major companies in Indian financial markets which includes:
    1. Bharti Airtel
    2. Reliance Industries
    3. Tata Steel
    4. Mahindra & Mahindra
    5. Tata Motors
    6. Tata Consultancy Services
    7. SBI
2. **Linguistic Inquiry and Word Count (LIWC):** The LIWC program includes the text analysis module along with a group of built-in dictionaries. LIWC reads written or transcribed verbal texts which have been stored in a digital, computer-readable form (such as text files). The text analysis module then then compares each word in the text against a user-defined dictionary. After the processing module has read and accounted for all words in a given text, it calculates the percentage of total words that match each of the dictionary categories. For example, if LIWC analyzed a single speech that was 2,000 words and compared them to the built-in LIWC2015 dictionary, it might

find that there were 150 pronouns and 84 positive emotion words used. It would convert these numbers to percentages, 7.5% pronouns and 4.2% positive emotion words.

3. **Data Imputation Technique:** The data imputation procedure implemented is described below:



Close	WC	Analytic	Clout	Authentic
36.759998	472	97.48	48.31	21.73
36.369999	279.6666667	96.46222222	51.66888889	64.51333333
35.939999	521.5	97.595	57.28	26.685
35.869999	313.5	95.4425	50.66	24.435
34.5				
34.869999	347.5	97.845	53.195	64.025
35.02	717	83.865	69.855	28.22
35.43	422	94.31	55.65	20.77
36.060001	153.5	94.465	77.65	10.84

Close	WC	Analytic	Clout	Authentic
34.98	505	99	52.38	64.13
35.790001	192.4166667	98.15333333	61.49916667	64.58416667
36.959999	12	99	50	74.76
36.150002	600.5	97.69	59.635	16.67
36.889999	195	99	62.1	36.78
35.889999	147	99	58.08	29.03

Close	WC	Analytic	Clout	Authentic
36.759998	472	97.48	48.31	21.73
36.369999	279.6666667	96.46222222	51.66888889	64.51333333
35.939999	521.5	97.595	57.28	26.685
35.869999	313.5	95.4425	50.66	24.435
34.5	399.7685582	95.66753369	59.73513229	47.62413713
34.869999	347.5	97.845	53.195	64.025
35.02	717	83.865	69.855	28.22
35.43	422	94.31	55.65	20.77
36.060001	153.5	94.465	77.65	10.84

It is a known fact that news articles corresponding to a particular stock doesn't appear everyday. This creates some situation of missing data in the dataset which can be filled logically without any specific imputation procedure. Stock markets remain closed on weekends and public holidays i.e, the stock prices are not available on these particular days while it may be possible that news articles may appear on these days. The information obtained from textual data on "missing days" (when stock price is missing) is then accomodated into the next instance when a new stock price appears after a set of "missing days" by averaging the LIWC scores of all "missing days" along with the scores on the day when stock price appears.

4. **Feature Selection Techniques:**
    1. **Chi-Squared Feature Selection**: Chi-squared test is used to test independence of two events or in other words, chi-squared test values help us to decide whether the predictor variable and the target class are independent. High chi-

squared values indicate the dependence of target variable on the predictor variable.

2. **MRMR Feature Selection: M**RMR or minimum redundancy maximum relevance feature selection method uses a heuristic framework to minimize redundancy, and a series of intuitive measures of relevance and redundancy to select promising features for both continuous and discrete data sets. The maximum relevance condition is obtained by selecting features using F-statistic values, to get the relevance scores for each predictor variable with respect to the target variable. In feature selection, it has been recognized that the combinations of individually good features do not necessarily lead to good regression performance. In other words, the best features obtained using maximum relevance condition may not combine to form the best feature set for the regression problem at hand. This generates the need for another step implementing the condition for minimum redundancy among the predictor variables. This is implemented in continuous variables using several techniques including correlation among predictor variables using Pearson coefficient, distance measures like Euclidian distance and many more.

5. **Machine Learning Techniques:**

1. **SVR:** A Support Vector Machine (SVM) performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories. SVM models are closely related to neural networks. Using a kernel function, SVMs are an alternative training method for polynomial, Radial Basis Function (RBF) networks and MLP classifiers, in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem, as in standard neural network training. The goal of SVM modeling is to find the optimal hyper plane that separates samples, in such a way that the samples with one category of the target variable should be on one side of the plane and the samples with the other category are on the other side of the plane. The samples near the hyper plane are the support vectors. An SVM analysis finds the hyperplane that is oriented so that the margin between the support vectors is maximized. One idea is that performance

on test cases will be good if we choose the separating hyper plane that has the largest margin.

2. **Random Forest (RF):** An ensemble with a set of decision trees is grown in randomly selected subspaces of data and average of their prediction is used for prediction purposes. Ensembles are a divide-and-conquer approach used to improve performance. Ensemble methods work on the principal that a group of weak learners can come together to form a strong learner. A small group of predictor variables is selected for each node to split on and the best split is calculated based on these predictor variables in the training set. The tree is grown using CART methodology to maximize size, without pruning. This randomization of subspace is combined with bagging to resample, with replacement, the training data set each time a new individual tree is grown. When a new input is entered into the system, it is run down all of the trees in the ensemble grown. The prediction result may either be an average or weighted average of all of the terminal nodes that are reached.

3. **Quantile Regression:** Quantile regression is a type of regression analysis used in statistics and econometrics. Whereas the method of least squares results in estimates that approximate the conditional *mean* of the response variable given certain values of the predictor variables, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. Quantile regression is desired if conditional quantile functions are of interest. One advantage of quantile regression, relative to the ordinary least squares regression, is that the quantile regression estimates are more robust against outliers in the response measurements.

4. **MLP:** A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

5. **GMDH:** Family of inductive algorithms for computer-based mathematical modelling of multi-parametric datasets that features fully automatic structural and parametric optimization of models. GMDH algorithms are characterized by inductive procedure that performs sorting-out of gradually complicated polynomial models and selecting the best solution by means of the so-called external criterion.

6. **RPART:** Recursive Partitioning and Regression Tree follows along the lines of close implementation of Classification and Regression Tree using the method called Automatic Interaction Detection (AID) performing stepwise splitting.

7. **Quantile Regression Random Forest (QRRF):** Quantile Regression Random Forest was introduced by Meinshausen. The key difference between the random forest and quantile random forest is that in random forest for each node in each tree, it keeps only the mean of the observations that fall into this node and neglects all other information. In contrast, quantile regression forests keep the value of all observations in this node, and assesses the conditional distribution based on this information.

8. **GRNN:** A GRNN is a variation of the radial basis neural networks, which is based on kernel regression networks. A GRNN does not require an iterative training procedure as back propagation networks. It approximates any arbitrary function between input and output vectors, drawing the function estimate directly from the training data. In addition, it is consistent that as the training set size becomes large, the estimation error approaches zero, with only mild restrictions on the function. A GRNN consists of four layers: input layer, pattern layer, summation layer and output layer as shown in. The number of input units in input layer depends on the total number of the observation parameters. The first layer is connected to the pattern layer and in this layer each neuron presents a training pattern and its output. The pattern layer is connected to the summation layer. The summation layer has two different types of summation, which are a single division unit and summation units. The summation and output layer together perform a normalization of output set. In training of network, radial basis and linear activation functions are used in hidden and output layers. Each pattern layer unit is connected to the two neurons in the summation layer, S and D summation neurons. S summation

neuron computes the sum of weighted responses of the pattern layer. On the other hand, D summation neuron is used to calculate un-weighted outputs of pattern neurons. The output layer merely divides the output of each S-summation neuron by that of each D-summation neuron, yielding the predicted value.

5.Results

The results are obtained on 7 datasets corresponding to 7 major companies including Bharti Airtel, Reliance Industries, Tata Steel, Tata motors, Mahindra and Mahindra, SBI, Tata Consultancy Services by application of 7 machine learning techniques involving GMDH, GRNN, RandomForest (RF), QRRF, RPART, SVR and MLP. The results were compared both with and without the application of feature selection techniques. The feature selection techniques employed were Chi-squared and Minimum Redundancy and Maximum Relevance (MRMR) feature selection techniques. The results are reported using MAPE and NRMSE statistical error measures.

MAPE = Mean Absolute Percentage Error:

$$\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y^{i}_{actual}-y^{i}_{predicted}}{y^{i}_{actual}}\right|\%$$

, where n = total number of observations.

NRMSE = Normalized Root Mean Squared Error:

$$\frac{1}{range}\sqrt{\frac{\sum_{i=1}^{n}\left(y^{i}_{actual}-y^{i}_{predicted}\right)^{2}}{n}}$$

, where n = total number of observations, range = maximum (observation set) – minimum (observation set)

Table 1. Stock Prediction Results (All features)

Dataset	GMDH		GRNN		RF		QRRF		RPART		SVR		MLP	
	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE
Airtel	0.120	0.018	1.467	0.238	7.12	0.614	6.12	0.469	17.41	0.708	4.70	0.56	3.14	0.43
Mahindra	0.524	0.047	2.75	0.175	16.9	0.594	16.54	0.478	17.9	0.72	15.6	0.57	5.83	0.26
Tata Motors	0.306	0.022	1.846	0.149	6.59	0.528	4.71	0.444	6.17	0.62	5.63	0.51	6.8	0.51
Reliance Industries	0.166	0.07	0.395	0.075	2.57	0.44	1.83	0.364	2.73	0.536	2.10	0.388	1.7	0.33
Tata Steel	0.616	0.041	1.268	0.129	12.58	0.774	9.74	0.67	17.01	0.994	10.15	0.74	7.2	0.64
TCS	0.114	0.025	0.179	0.04	3.02	0.476	2.50	0.408	3.08	0.596	2.92	0.454	1.9	0.4
SBI	0.494	0.040	3.30	0.26	4.98	0.624	3.02	0.532	7.16	0.836	4.81	0.568	7.3	0.51

Table 2. Stock Prediction Results (Chi-square-top-25)

Dataset	GMDH		GRNN		RF		QRRF		RPART		SVR		MLP	
	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE
Airtel @	1.09	0.166	1.75	0.26	4.94	0.61	4.07	0.48	5.85	0.74	4.69	1.09	3.42	0.366
Mahindra *														
Tata Motors *														
Reliance Industries *														
Tata Steel	2.85	0.275	2.55	0.261	10.74	0.76	7.48	0.58	12.73	0.97	9.93	0.76	7.64	0.597
TCS	0.270	0.056	0.395	0.118	2.2	0.47	1.78	0.402	2.89	0.65	1.98	0.42	1.37	0.297
SBI #	2.007	0.163	4.49	0.36	9.03	0.654	8.1	0.55	10.44	0.80	8.11	0.612	8.33	0.545

Table 3. Stock Prediction Results (Chi-square-top-10)

Dataset	GMDH		GRNN		RF		QRRF		RPART		SVR		MLP	
	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE
Airtel	1.49	0.230	2.831	0.366	5.20	0.65	3.93	0.50	6.44	0.84	4.94	0.59	3.95	0.495
Mahindra	3.60	0.317	4.37	0.347	7.33	0.52	6.03	0.45	8.44	0.65	8.65	0.63	6.281	0.068
Tata Motors	4.60	0.346	4.67	0.346	7.3	0.50	5.9	0.41	8.64	0.622	7.58	0.52	6.49	0.449
Reliance Industries	1.45	0.227	1.391	0.279	2.84	0.44	2.3	0.36	3.44	0.55	2.78	0.44	2.607	0.407
Tata Steel	4.32	0.365	5.30	0.425	10.36	0.76	8.3	0.68	11.41	0.87	10.35	0.77	7.44	0.561
TCS	1.14	0.240	0.902	0.217	2.172	0.447	1.71	0.37	2.57	0.56	2.24	0.44	1.851	0.366
SBI	3.60	0.29	5.25	0.402	9.067	0.649	7.91	0.565	11.06	0.81	8.26	0.603	7.99	0.509

Table 4. Stock Prediction Results (Stock Prediction Results (MRMR-top-25)

Dataset	GMDH		GRNN		RF		QRRF		RPART		SVR		MLP	
	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE
Airtel	0.325	0.048	1.385	0.177	5.128	0.624	3.631	0.437	5.39	0.663	4.67	0.567	3.84	0.0453
Mahindra	1.49	0.133	2.33	0.172	7.52	0.546	6.60	0.474	9.045	0.659	8.08	0.556	6.209	0.468
Tata Motors	1.58	0.131	1.81	0.148	7.21	0.506	6.71	0.457	8.257	0.597	7.278	0.508	5.97	0.449
Reliance Industries	0.608	0.097	0.568	0.110	2.534	0.419	2.28	0.377	3.37	0.552	2.231	0.372	2.067	0.334
Tata Steel	2.04	0.185	1.62	0.151	10.25	0.736	7.39	0.643	12.203	0.927	9.23	0.718	6.73	0.515
TCS	0.309	0.067	0.319	0.079	2.123	0.450	1.83	0.388	2.821	0.6001	2.104	0.441	1.69	0.320
SBI	1.82	0.154	4.49	0.360	8.67	0.612	7.364	0.518	11.13	0.820	7.42	0.574	7.76	0.518

Table 5. Stock Prediction Results (MRMR-top-10)

Dataset	GMDH		GRNN		RF		QRRF		RPART		SVR		MLP	
	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE	MAPE	NRMSE
Airtel	1.35	0.207	2.34	0.308	4.98	0.620	4.145	0.482	5.361	0.708	5.211	0.647	4.303	0.541
Mahindra	3.27	0.275	4.37	0.316	7.65	0.561	6.87	0.489	8.702	0.662	8.35	0.601	5.94	0.461
Tata Motors	3.61	0.274	3.62	1.39	6.96	0.487	5.975	0.431	7.74	0.566	7.19	0.503	5.88	0.42
Reliance Industries	1.232	0.206	1.116	0.203	2.51	0.415	2.195	0.358	3.34	0.550	2.158	0.366	1.91	0.31
Tata Steel	3.81	0.347	4.71	0.418	9.34	0.679	6.773	0.568	11.85	0.895	9.204	0.692	6.42	0.513
TCS	0.929	0.209	0.863	0.209	2.02	0.438	1.875	0.380	2.78	0.572	1.715	0.387	1.46	0.306
SBI	3.60	0.29	5.25	0.402	8.47	0.598	6.91	0.512	10.253	0.759	7.64	0.576	7.37	0.50

The above tables describe the results for the stock price prediction analysis for the extracted datasets. The implemented models include GMDH, GRNN, RandomForest, QRRF, RPART, SVR and MLP. The results clearly show that the GMDH performs best on all the datasets followed by the GRNN and MLP. Comparable results were obtained with QRRF and Random Forest. One of the reasons for this observation can be that GMDH is cited as one of the earliest Deep learning algorithms in the data mining paradigm which justifies its supremacy in this case over other approaches. Neural Networks are seen to perform better than other algorithms which are clearly evident from the fact that GRNN and MLP occupies second and third place in terms of the goodness of their accuracy in stock prediction.

6.Conclusions

In this paper, we proposed the model for stock market prediction based on published news articles with respect to a stock (company). We employed various models including RandomForest, QRRF, GMDH, SVR, RPART, MLP and GRNN for regression. We conducted experiments on 7 major companies in the Indian stock markets. We performed the 5-fold Cross-validation method to evaluate the model performance. In our experiments GMDH is giving best performance compare to other model followed by GRNN.