Stock Price Prediction using Mean Level Corporate Communication Frequency

Gowtham Sethupathi M, Chaitanya P, Hariprasath R, Rathyin V R, Suresh Kumar M

Abstract*: Researches in corporate communication network have revealed existence of certain patterns and these patterns can potentially reveal useful information about the organization. Different communication patterns have been formed among different employees of an organization; in this paper we analyze the communication network of an organization to observe the underlying human social behavior that is expressed in an email system. It is a known fact that email systems form a social network. The overall communication pattern in an organization can be directly attributed to its stability and performance; and the stock market value of that organization can be predicted based on the different communication patterns which are obtained from the analysis. These patterns have proved to be successful in helping predict sales value and stock prices. Investing in share market requires lots of study and analysis; and also has certain risks. We use Ensemble model which consists of different prediction models such as ARIMA (Auto Regression Integrated Moving Average), LSTM (Long-Short Term Memory), etc. to predict the stock movements in this model. As Enron Corporations e-mail database is the only corpus publicly available, we use it as the data set for our data mining algorithms.

Index Terms*: ARIMA, Communication pattern, Ensemble method, Stock market.

I. INTRODUCTION

Communication plays an important role in increasing the efficiency of employees; they need to interact with each other on a regular basis to form a good rapport and be comfortable with working together. Problems arise when the flow of information in the desired direction is interrupted in the workplace. Effective communication facilitates free exchange of information involving the employees and minimizes misunderstandings and confusions. It ensures that everybody is on the same level; also it plays a compelling role in motivating workers. Managers must move with their team members on an everyday basis for them to develop a perception of loyalty towards the organization. Team members must be made to feel important. Managers must sit with them, provide correct feedback, guide them and facilitate them to trot out changes with a smile. Employee performance increases once their grievances are properly addressed by their reporting managers. Workplace communication contributes to increase in productivity. Research states proven facts that effective communication between work groups results in enhancement of company-wide performance. On top of this; it has been proven that employees holding highest rankings in productivity have been maintaing a great communication rapport with their superiors. For example John works in software engineering and his prototypes of applications receive positive comments from everyone. He effectively communicates about his algorithms and the way they're going to be used in reality by clients. In addition He takes specific inputs and suggestions from his superiors allowing him to create software products swiftly and with efficiency.

Workplace communication can potentially increase employee job satisfaction. Employees feel capacitated if they are capable of communicating upwards in the chain of command. In such a form of communication; data flows upwards in a company and habitually consists of feedback. If higher officers are capable of listening and responding to employees, it leads to an increase in employee job satisfaction. Moreover, employees are also in good spirits if there is a keen downward communication. Example, John sends an email to his team leader recommending that the team make certain upgrades to its designing software. His suggestions in the upward communication were not only recognized but also exercised, resulting in an upbeat design team. Recent study reveals the existence of interesting connections between communication patterns among different participants of an organization and organizational stability in a corporate network. These patterns have proved to be useful in predicting sales values and stock prices. A corporate network connects employees within a big corporation, and the members of a corporate communication network mainly exchange company- specific business information. E-mails have been employed as a means for inter-organizational and intra-organizational information exchange for a long time in a corporate communication network. In line with corporate communication theory, which suggests that employee communications can make or break any major change arising from a merger, new venture, acquisition of child companies, or other management issues; employee communication can possibly serve a critical business function that drives performance and contributes to a company's financial success.

Revised Manuscript Received on 30 May 2019.

* Correspondence Author

Gowtham Sethupathi M*, Department of CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Chaitanya P, Department of CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Hariprasath R, Department of CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India..

Rathyin V R, Department of CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Suresh Kumar M, Department of CSE, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Given the increase in usage of communication networks, the mining of interesting communication patterns could provide insights into the development of many useful applications in other various possible cases. Enron Corporations e-mail database is the only corpus publicly available; we use it as the data set for our stock prediction system. The Enron corpus has already been classified and tagged with keywords in order to be easily used for research purposes. Thus; the raw data can be dispatched without any preprocessing into the algorithm. To imply the prediction system upon other datasets; a natural language processing module can be employed; which would be used to pre-process the e-mail corpus to categorize the data into useful classes that include messages containing official and financial data. The rest of the mail that is unnecessary such as invitations and general management instructions can be filtered out discarded. This allows for the effective use of this stock prediction system to be used in many other practical applications.

II. LITERATURE SURVEY

A. Data Mining Of Financial News for Major Events and Their Effects on the Market [2]

Submit your manuscript electronically for review. Analysis of the financial market predominantly involves a huge amount of information in the form of structured data which reflects the economic performance of the recipient; text mining can be used to obtain information about relevant events from stock market news. It also explains the causes for erratic fluctuations in the market. Recently text mining is being used for the prediction of stock price of individual companies, Information collected from eclectic sources are combined and used to construct strategies that assist potential investors. This paper uses a text mining system to analyze the Indian stock market based on finance market news and correlates it with the actual behavior of the stock market. The aim is to identify the noteworthy events that have had a considerable impact on the stock market and characterize them in order to construct prediction rules for predicting the market. Financial time series data for Bombay Stock Exchange for three years (2005-2008), and chronologically ordered financial articles for the same time period obtained online are used as the dataset for this analysis. However these systems failed to identify many other factors that affect the stock market as a whole This paper uses deep learning methods such as Recursive Neural networks RNN for prediction of stock market movements using financial news articles and a set of technical indicators as input Gradient vanishing and exploding problems are the major disadvantages of RNN it makes the training of datasets difficult

B. Stock Market Prediction using Data Mining Techniques [3]

In this model; Natural Language Processing and data mining are used to process news articles and other relevant sources. Ensemble learning models such as random forest and support vector machine are used for processing the data. Classification is used to solve the time series forecasting problem; which is regarded as an example of supervised learning. In supervised learning method; the dataset containing the observations is used for learning by training a dataset with previously identified classes. Random Forest model is used; which creates multiple decision trees based on varying subsets of data during the training and outputs the mode of classes that have resulted from these decision trees using an input for classification. Elongated periods of training time for large datasets, Difficulty in understanding and interpreting the final model, and inability to make small calibrations due to variable weights and individual impact are major disadvantages with this system.

C. Stock Market Prediction Using Machine Learning Approach [4]

In this paper well-known efficient regression approaches such as multiple regression, linear regression, and Polynomial regression are surveyed to predict the stock price from stock market data. Major concerns of this system are that linear regression only looks at linear relationships between dependent variables and independent variables. A linear relationship is assumed between them which are incorrect sometimes. For example, the correlation between age and income is nonlinear, i.e., income tends to rise in the early stages of adulthood, become flat in later adulthood; and then declines after retirement. Demerits of using a multiple regression model usually boils down to the data that is used being incomplete and basely consummating that an association is causation. The authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission.

D. Optimized Stock Market Prediction using Ensemble learning [5]

The financial forecasting problem is approached by training an ensemble of classification models on a set of financial data. This set of data is categorized into sets of ups and downs based on whether there was an upward or downward inclination in stock value during the observed period of time. The learning model is designed in such a way that top priority is given to the aspect of feature selection; and also that it is a fully automated procedure. This allows for the model to decide on its own on which parameters to choose to make an effective prediction instead of being strained to accept variables that are assigned by humans. This gives the ensemble method an additional feature to be flexible meaning that now it can make up its own formulations and attributes according to need.

E. Technical Communication for Organizational Knowledge Creation [6]

Organizational knowledge is studied which reveals information crucial to the functioning of an organization and its performance. This paper describes the process of creation of organizational information and the role of technical writing and; professional communication. After constructing the knowledge creation process in a discrete time Markov chain form, occupancy time is presented and evaluated to show the role it plays in affecting organizational performance.

Numerical examples have also shown that technical writing and professional communication can provide valuable information and hence can act as an enabler for organizational growth. Despite the knowledge of such potential importance about corporate communication, only a little amount of work has been done in this important direction to include them in data mining techniques. Further studies and experiments can be used to arrive at a conclusive point; at which more effort and time can be invested in this research in this direction.

F. Effect of Corporate Communication on Stock Price Movements [7]

This work anchors on finding if a relation exists between a corporate communication network and its financial productivity and successfully demonstrates that a corporate e- mail system contains meaningful information about employees' communication patterns. A corporate company has identifiable patterns of e-mail exchange which can reveal important information about its activities such as mergers, business deals, acquisitions etc. and organizational stability that might subsequently influence the particular company's performance in the market. Therefore, corporate communication patterns can act as a valid indicator to predict a company's stock movements. The findings in this study gives scope to a great extent to further implement upgrades and as a result; increase accuracy levels in the process of prediction of stock market movements.

III. SYSTEM ARCHITECTURE

The system architecture of the proposed system is depicted in Fig.1. The architecture is basically split into five modules which are:

- Dataset Module
- Weight Matrix Construction
- Discretization of Values
- Communication Pattern Recognition
- Prediction

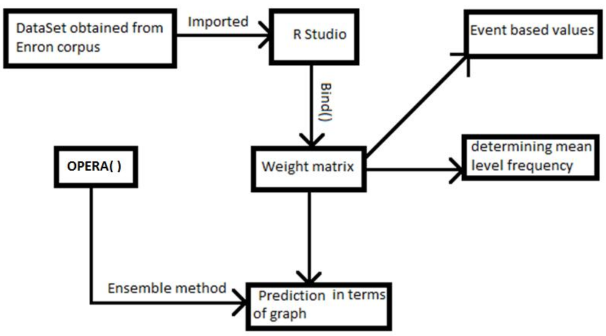

Figure 1: System Architecture

The dataset module consists of the Enron e-mail corpus which is downloaded from the internet and is imported into R studio using the bind ( ) function. The second module is weight matrix construction, here the communication frequency [15] of the employees in the corporation is mapped to a graph and then they are stored in the form of matrices. The third module is discretization of values, Continuous stock data and communication data are stored as discrete values. Example of an event is rise of fall in the value of stock price. The fourth module is communication pattern recognition. Evidence weights are assigned to every communication based on the effect they have on stock value. Total weight of evidence is calculated to determine the influence of communication frequency upon stock movements. The final module is the prediction module. R packages such as OPERA are used to combine different prediction algorithms like ARIMA [13] and LSTM and finally results are plotted.

IV. METHODOLOGY

We predict the stock price movements of an organization, by analyzing the organizations communication along with its stock value. The proposed algorithm includes the following steps:

- Obtaining dataset
- Weight matrix construction
- Discretizing weight matrix values and recognizing communication patterns
- Construction of prediction rules

A. Obtaining Dataset

Enron organization has made their corpus dataset public for research purposes. It contains authentic e- mail messages that were exchanged within the corporation's employees. It contains 96.3 million words and a rough estimate of (half a million) 500000 number of e-mails. "Kaggle" website is used to obtain the e-mail dataset from the organization. It contains pre classified email into user distinctive folders and information obtained from sequentially arranged email threads. This dataset has been downloaded in the .csv format. Each cell is separated by a comma, other characters can be used. Once the dataset has been downloaded, it is imported into R Studio using CSV File package. The command 'mydata = read.csv (mydata.csv)' reads the csv file and imports it into R Studio which is then converted into a spreadsheet. Since it is a csv file; the value for each message should be initialized for it to be processed properly. The imported dataset is stored in the name of mydata. Based on the classification [8] of emails, unwanted e-mails are removed from the csv file. This is done by selecting the spam messages from the classification, messages which have been sent more than one time will be removed from the dataset. A separate table is created for storing employees' weight matrix. The table is named as matrixdata.

Two columns are created; one consists of employee id and other consists of the matrix data which needs to be constructed. Two columns are initialized for storing values. Once it is done; the weight matrix needs to be constructed.

As the Enron corpus data that is being used in this prediction system has been pre-processed to eliminate unwanted e-mails; it poses no issues; but for future usage in other datasets, the input data needs to be analyzed and classified into useful and useless.

To achieve this; NLTK (Natural Language Processing toolkit) [9] is used. Initially, the text messages are acquired from the corpora available in NLTK (a suite available in Python), and the corpus is analyzed. Next classification of hierarchical topics is done using Fuzzy Logic, followed by unsupervised learning method [10]. The processed dataset is then stored in a csv file format in python and is then easily transferred to R studio platform.

B. Weight Matrix Construction

The data that represents the communication frequency between the employees is denoted in a graph form containing 2 tuples $G=(V_t, R_t)$ where:

$V_t = \{v1, v2, \ldots vn\}$ represents a finite set of nodes (employees in the communication network), n is the total number of nodes present in the network, and v_c represents the central node of the network. In this case, we choose the CEO as the central node.

$R_t = \{vij\ t,\}$ denotes the set of edges representing all the links in the network, where $R \subseteq V \times V$, where $v_{ij,t} \in [L_{Vj}, U_{Vj}]$, L_{Vj} represent the lower bound and U_{Vj} represents upper bound of the values, $1 \le \tau \le t - 1$ $\tau \in Z_+$ and $t = 1, 2, \ldots p$, where p is the total number of time points.

To gain better perception of the data, Fig.2 Is used to describe the input in graph format. Six nodes in the graph represent six employees in the communication network, respectively. The edge of two nodes denotes the communication frequency between two workers. The whole input data describes the changes of communication network during five time points.The graphical data is converted into an n X n matrix $[d_{ij}]$ $(1 \le i, j \le n)$ so that the value of d_{ij} can capture the frequency of communication between any given nodes within the corporate communication network. This matrix is constructed to facilitate the detection of communication patterns; as shown in Fig.3.

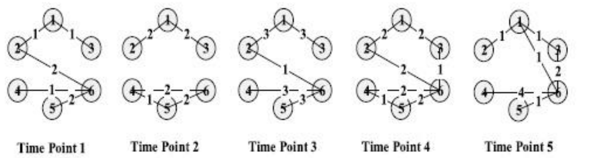

Figure 2: Communication Frequency

$$\begin{pmatrix}0&1&1&0&0&0\\1&0&0&0&0&2\\1&0&0&0&0&0\\0&0&0&0&0&1\\0&0&0&0&0&2\\0&2&0&1&2&0\end{pmatrix}
\begin{pmatrix}0&2&3&0&0&0\\2&0&0&0&0&0\\3&0&0&0&0&0\\0&0&0&0&1&2\\0&0&0&1&0&2\\0&0&0&2&2&0\end{pmatrix}
\begin{pmatrix}0&3&3&0&0&0\\3&0&0&0&0&1\\3&0&0&0&0&0\\0&0&0&0&0&3\\0&0&0&0&0&3\\0&1&0&3&3&0\end{pmatrix}
\begin{pmatrix}0&2&2&0&0&0\\2&0&0&0&0&2\\2&0&0&0&0&1\\0&0&0&0&1&2\\0&0&0&1&0&2\\0&2&1&2&2&0\end{pmatrix}
\begin{pmatrix}0&1&1&0&0&1\\1&0&0&0&0&0\\1&0&0&0&0&2\\0&0&0&0&0&4\\0&0&0&0&0&1\\1&0&2&4&1&0\end{pmatrix}$$

Figure 3: Weighted Matrix

C. Discretizing values and Recognizing Communication Patterns

Data discretization converts a plethora of values into smaller ones, so that management and evaluation of data becomes very easy. In this section, we discuss about the discretization [10] of values and recognition of communication patterns. The sections are organized as follows: C-1 explains the discretization of weighed communication matrix; C-2 discusses the steps involved in discretizing the stock movements; and C-3 elaborates the pattern recognition process.

1. In this case, we take into account that the habits of exchanging e-mail messages varies in each person based on their personal behavior and also the roles they play in a organization. Example of the above mentioned scenario; a secretary needs to send a plethora of e-mails for coordination in daily course of work. Therefore, the mean level; for instance 20 e-mails per day, is considered as the normal number for a secretary according to their communication pattern. Relatively; 20 e-mails per day might be already well above the mean level of the communication frequency [15] for an employee of lower hierarchy. Thus, the communication level is standardized according to each person/node's mean of their communication frequency. To account for this variation in communication frequency; we use equal frequency algorithm [10] to discretize the values of weights in the communication network.

A set of event-based values are denoted as $Eij = \{Eij,1, Eij,2, \ldots Eij,t \ldots, Eij,p\}$, Eij,t _Eij correspond to $vij,t \in Rt$, where the domain of the value of Eij is denoted as dom(Eij) = [Lij, Uij], in which Lij represents the lower bound and Uij represents the upper bound of the values. Hence, the set of discrete states for Eij,t is denoted as $D(Eij,t) = \{dr1, dr2 \ldots drk \ldots drn$, where dr$k$ represents the discrete states for the value of $(Rt\ /GSt\ )$, n is the total number of states. All continuous data are remodeled into discrete data such as dr1, dr2, dr3. After successful discretization of the matrix, the levels of communication between employees are classified into 'Weak', 'Strong', and 'None'.

2. The stock movements of an organization are continuous in their raw form and need to be discretized. To achieve this; the original value of increases (up) and decreases (down) in the stock price are represented using the price fluctuation ratio; calculated by $(CPt{-}1{-}CPt\ /\ CP{-}1)$ where CPt−1 represents closing price of a stock on the day "t−1" and CPt represents the closing price on the day "t". Next, the movements of stock price are classified into discrete states, such as ds1, ds2, which, respectively, represents the ratio of increase of stock price is higher than 0 and the ratio of decrease of stock price is lower than 0. Finally, the original matrix is discretized as $Rt = \{Eij,t \mid Eij,t \in D(Lij,t), D(Lij,t) = \{dri\}, 1 \le i, j \le n\}$ and the value of movement of stock price is discretized as $st = Et$, where $Et = D(Lt)$, and $D(Lt) = \{dsi\}$.

Table 1: Frequency - Event based value Vs. Stock Price

o_{ij} (e_{ij})		Stock Up (ds_1)	Middle (ds_2)	Down (ds_3)	Total
$E^c_{j,t-1}$	None (dr_1)	0 (0)	0 (0)	0 (0)	0
	Strong (dr_2)	0 (0)	1 (0.5)	0 (0.5)	1
	Weak (dr_3)	0 (0)	1 (1.5)	2 (1.5)	3
Total		0	2	2	4

3. The associations between mean communication frequency and stock price movements that exhibit communication patterns are detected using residual analysis [11]. Let *P(Eij ,t − τ* = dr*i* → *Et* = ds*i)* be the graph stock pattern between *Eij,t− τ* = dr*i* and *Et* = ds*ij*. *Eij,t− τ* represents the e-mail communication frequency between any two nodes of the network, and *Et* represents the movement of stock price in the time points of *t*; *τ* is the time interval and 1 ≤ *τ* ≤ *t* − 1, *τ* ∈Z+. A two dimensional contingency of *J* rows and *I* columns is constructed as shown in Table I. to discover the underlying patterns. *I* is the total number of states of stock price movement, and *J* is the total number of states of relationship between the central node *(vc)* and jth node *(vj)*. In Table I, *oij* is the total number of counts when the value of communication between *vc* and *vj (Ecj,t−* 1*)* is dr*j* at *t*th time points, and the value of stock price movement (*Et)* is ds*i* after one time instants. *eij* is the summation of o*iu* (from u= 1-J) and o*u j* /M' (from u= 1-I); it is used to capture the expected total number under the assumption that *Et* and *Ecj t−* 1 are independent. If the value of *eij* is lower (negative), the value of stock price depends on the e-mail communication. On the contrary, if the value of *eij* is higher, they are more independent upon each other.

Table 2: 2D Contingency to discover graph stock pattern

o_{ik} (e_{ik})		Stock Price (in the certain position later) ds_1	…	ds_i	…	ds_I	Total
$E^c_{j,t-1}$	dr_1	o_{11} (e_{11})	…	o_{i1} (e_{i1})	…	o_{I1} (e_{I1})	o_{+1}
	⋮	⋮		⋮		⋮	⋮
	dr_j	o_{1j} (e_{1j})	…	o_{ij} (e_{ij})	…	o_{Ij} (e_{Ij})	o_{+j}
	⋮	⋮		⋮		⋮	⋮
	dr_J	o_{1J} (e_{1J})	…	o_{iJ} (e_{iJ})	…	o_{IJ} (e_{IJ})	o_{+J}
Total		o_{1+}	…	o_{i+}	…	o_{I+}	M'

D. Constructing the Prediction Module

The prediction rules are constructed based on weighted evidences [12]. The weight (W) of evidence is used to measure the amount of negative or positive influence it has over an attribute; which in this case is the interrelation between communication frequency and the stock price movements. It denotes whether the interrelation between the values agree or disagree with the movement of stock prices. A negative value of weighted evidence means that the communication pattern has no influence over the stock value; and a positive value indicates that the two are indeed related. The weight evidence measure for each association is defined in terms of mutual information as:

$$I(E_t = \mathrm{ds}_i : E^c_{j,t-1} = \mathrm{dr}_j) = \log \frac{Pr(E_t = \mathrm{ds}_i | E^c_{j,t-1} = \mathrm{dr}_j)}{Pr(E_t = \mathrm{ds}_i)} \quad (1)$$

All the weight values for the evidence are calculated using (2).

$$\begin{aligned} &W(E^c_{j,t-1} = dr_j \rightarrow E_t = ds_i) \\ &= I(E_t = \mathrm{ds}_i : E^c_{j,t-1} = \mathrm{dr}_j) - I(E_t \neq \mathrm{ds}_i : E^c_{j,t-1} = \mathrm{dr}_j) \\ &= \log \frac{\Pr(E_t = \mathrm{ds}_i | E^c_{j,t-1} = \mathrm{dr}_j)}{\Pr(E_t = \mathrm{ds}_i)} - \log \frac{\Pr(E_t \neq \mathrm{ds}_i | E^c_{j,t-1} = \mathrm{dr}_j)}{\Pr(E_t \neq \mathrm{ds}_i)} \\ &= \log \frac{\Pr(E_t = \mathrm{ds}_i, E^c_{j,t-1} = \mathrm{dr}_j)}{\Pr(E_t = \mathrm{ds}_i) \cdot \Pr(E^c_{j,t-1} = \mathrm{dr}_j)} - \log \frac{\Pr(E_t \neq \mathrm{ds}_i, E^c_{j,t-1} = \mathrm{dr}_j)}{\Pr(E_t \neq \mathrm{ds}_i) \cdot \Pr(E^c_{j,t-1} = \mathrm{dr}_j)} \\ &= \log \frac{\Pr(E_t = \mathrm{ds}_i, E^c_{j,t-1} = \mathrm{dr}_j) \Pr(E_t \neq \mathrm{ds}_i) \cdot \Pr(E^c_{j,t-1} = \mathrm{dr}_j)}{\Pr(E_t \neq \mathrm{ds}_i, E^c_{j,t-1} = \mathrm{dr}_j) \cdot \Pr(E_t = \mathrm{ds}_i) \cdot \Pr(E^c_{j,t-1} = \mathrm{dr}_j)} \\ &= \log \frac{\Pr(E_t = \mathrm{ds}_i, E^c_{j,t-1} = \mathrm{dr}_j) \Pr(E_t \neq \mathrm{ds}_i)}{\Pr(E_t \neq \mathrm{ds}_i, E^c_{j,t-1} = \mathrm{dr}_j) \cdot \Pr(E_t = \mathrm{ds}_i)} \end{aligned} \quad (2)$$

After individual weights have been calculated and assigned to events; they are combined together to generate Total Weight (TW); which ultimately shows the extent of dependency between communication pattern and stock movements. Total weight is calculated using (3).

$$\mathrm{TW}\left(\frac{E_t = \mathrm{ds}_i}{E_t \neq \mathrm{ds}_i} | v[1], v[2], \ldots v[m]\right) = \sum_{j=1}^{\beta} W\left(\left(\frac{E_t = \mathrm{ds}_i}{E_t \neq \mathrm{ds}_i}\right) | E^c_{j,t-1} = dr_j\right) \quad (3)$$

Finally, once the prediction rules are set; a small fraction of the data set is input as training data for the rules. Then the rest of the records as given as testing data input; this is the data upon which the actual prediction algorithm is executed upon; and the predictions are made. The precision of the prediction is calculated by a division between the corrected number of predictions in the testing data and the total count of testing data.

V. RESULTS AND DISCUSSION

The actual stock price values of Enron from March 2001 to December 2001 that is publicly available is used for testing the prediction system as the obtained email corpus corresponds to the respective timeline of stock price movements. An ensemble model combining various time series forecasting methods such as ARIMA, and LSTM (Long-Short Term Memory) is used for the prediction of the stock prices. Ensemble model works by implementing a number of algorithms on a dataset and then combines the results of individual forecasting models to obtain a combined prediction which is significantly higher in terms of accuracy. R packages are used to combine the algorithms; they are available online which can be downloaded and imported to the R studio. 'OPERA' (Online Prediction by ExpeRt Aggregation) and 'forecastHybrid' are R packages that are used for combining time series forecast algorithms together.

The 'forecastHybrid' model combines multiple models by clubbing together the values whose weights are equal.
The OPERA package operates by providing expert forecasts based on previous performances that are obtained from the internet in the R user crowd source.

A week is taken as the time interval; which means that the mean communication frequency of the previous week is monitored and the stock movements for the next week are predicted from this observation. Two level training system is used; them being UP and DOWN. Up indicates the percentage of increase in stock value is greater than zero and DOWN indicated that the percentage of decrease in stock value is lesser than zero. Table 3 shows the prediction result for the above mentioned training and testing parameters. The predicted stock price graph is shown in Fig.4.

Table 3: Experimental Results

Time interval	Training data	Testing data	Training accuracy (Two level)
One Week	90	45	73.33%
	95	40	80%
	100	35	76.62%
	105	30	86.67%
	110	25	83.3%
Average Accuracy			80.184%

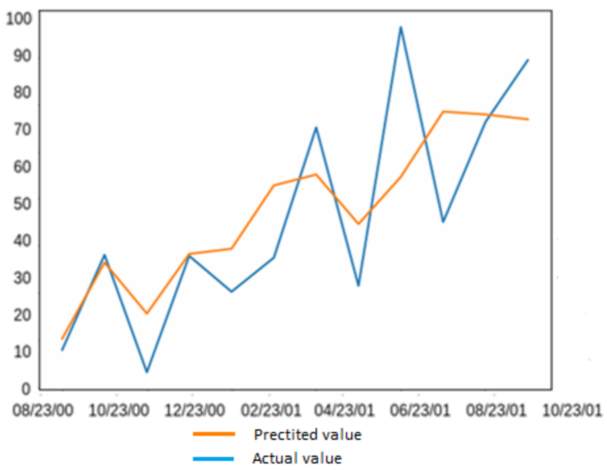

Figure 4: Actual vs. Predicted Stock values

VI. CONCLUSION AND FUTURE WORK

In summary, this paper proposes a stock price prediction system based on communication frequency of an organization. Ensemble method is used to combine different forecasting models to increase the accuracy of the prediction system. This eliminates the risks of investment in the stock market. This proposed system has achieved an accuracy level of approximately 80 percentage. From this we can inversely deduce that the stock value of an organization can be increased with better communication. Organizational stability can be achieved by this model. It can also be used to alter communication strategies in an organization with the help of pattern recognition. This paper paves way for future developments such as implementing fuzzy data sets instead of discrete data. NLP has been used to categorize useful and useless data only to monitor communication frequency, but it could further be utilized to implement text mining and understand the context of messages and improve the accuracy of the algorithm to a greater extent.

REFERENCES

1. Lailei Huang, Zhengyou Xia, "User Character and Communication Pattern Detecting on Social Network Site", 2009 International Conference on Engineering Computation.
2. Manuel R. Vargas, Beatriz S. L. P. de Lima, Alexandre G. Evsukoff, "Deep learning for stock market prediction from financial news articles", 2017 IEEE International Conference on CIVEMSA.
3. Sahaj Singh Maini, Govinda.K, "Stock Market Prediction using Data Mining Techniques", 2017 International Conference on Intelligent Sustainable Systems (ICISS).
4. Ashish sharma, Dinesh bhuriya, Upendra singh, **"**Survey of Stock Market Prediction Using Machine Learning Approach"**,** International Conference on Electronics, Communication and Aerospace Technology (ICECA) 2017.
5. Muhammad Asad, "Optimized stock market prediction using ensemble Learning", 2015 9th International Conference on Application of Information and Communication Technologies (AICT).
6. Takashi Okuda**, "**Technical Communication for Organizational Knowledge Creation"**,** 2001, International Professional Communication Conference.
7. Pei-Yuan Zhou, Keith C.C. Chan, Carol Xiaojuan Ou, "Corporate Communication Network and Stock Price Movements: Insights From Data Mining", IEEE Transactions on Computational Social Systems.
8. Bryan Klimt, Yiming Yang, "The Enron Corpus: A New Dataset for Email Classification Research", 2004, European Conference on Machine Learning: pp. 217-226.
9. S.Seshathri Aathithyan, M.V.Sriram, S.Prasanna, Dr.R.Venkatesan, "Affective – Hierarchical Classification of Text – An Approach Using NLP Toolkit", 2016, [ICCPCT].
10. Andrew K. C. Wong, David C. C. Wang, "DECA: A Discrete-Valued Data Clustering Algorithm", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. PAMI-1, NO. 4, October 1979.
11. Shelby J. Haberman,"The Analysis of Residuals in Cross-Classified Tables" International Biometric Society, Vol. 29, No. 1 (Mar 1973), pp. 205-220.
12. Yang Wang, Andrew K.C. Wong, "From Association to Classification: Inference Using Weight of Evidence", IEEE Transactions on Knowledge And Data Engineering, Vol. 15, No. 3, May/June 2003.
13. Ayodele A. Adebiyi., Aderemi O. Adewumi, Charles K. Ayo, "Stock Price Prediction Using the ARIMA Model", 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.
14. Li-Chen Cheng, Yu-Hsiang Huang, Mu-En Wu, "Applied attention-based LSTM neural networks in stock prediction", 2018 IEEE International Conference on Big Data.
15. Masaki Aida, Yudai Honma, Chisa Takano, Hideyuki Koto, "Unified Model of Communication Frequency between Users in Social Networks", 2010 IEEE International Conference on Communications.

AUTHORS PROFILE

Gowtham Sethupathi M, Received M.E. degree from Anna University in the year 2015. His research interests include Natural Language Processing. He is an Assistant Professor (O.G) for the Department of Computer Science and Engineering, SRM Institute of Science and Technology. Ramapuram Campus.

Chaitanya P, pursuing B. Tech degree in the department of Computer Science and Engineering, Ramapuram Campus, graduating in the year 2019. His research interests include data mining, Internet of Things and embedded systems.

Hariprasath R, pursuing B. Tech degree in the department of Computer Science and Engineering, Ramapuram Campus, graduating in the year 2019. His research interests include data mining, Internet of Things and data analytics.

Rathyin V R, pursuing B. Tech degree in the department of Computer Science and Engineering, Ramapuram Campus, graduating in the year 2019. His research interests include data mining, Internet of Things and data analytics.

Suresh Kumar M, pursuing B. Tech degree in the department of Computer Science and Engineering, Ramapuram Campus, graduating in the year 2019. His research interests include data mining, Internet of Things and Natural Language Processing.