Prediction of Box Office Success of Movies using Hype Analysis of Twitter Data

Sameer Thigale, Tushar Prasad, Ustat Kaur Makhija, Vibha Ravichandran

Abstract*— *Internet and Social Networking play a vital role in research field. It contains a massive diction about what people think. Twitter, is a micro blogging site where people post their views and preferences related to their interests. In this project, we try to predict the box office success of the movie by analyzing the hype created amongst the mob. We use sentiment analysis of Twitter data for the same. We are also considering the distribution area of the movie along with its genre. To display the output we plot the graph which depicts the success ratio of the movie.

Keywords—prediction; social networking; regression; sentiment analysis;

I. INTRODUCTION

Social networking is used by majority of the population around the world. Communication has been simplified due to social media. In digital age where speed is the main aspect to be followed social networks play a vital role in connecting people living across the globe.

Since Social Media is also a huge collection of wisdom we discovered the fact that this knowledge can be used in a positive manner to make predictions. Using social media to predict the future has become very popular in recent years

People have tried to show that twitter-based prediction of box office revenue performs better than market based prediction by analyzing various aspects of tweets sent during the movie release. The Twitter and YouTube data is used to predict the IMDB scores of movies. Sentiment analysis of twitter data is also in demand in recent years. There are also many other factors that affect the box office success of a movie. While sentiment analysis of documents has been studied for a long time, the techniques may not perform very well in twitter data because of the characteristics of tweets.

The following are a few difficulties in processing twitter data: the tweets are usually short (up to 140 words). The text of the tweets is often ungrammatical. It investigates features of sentiment analysis on tweets data. However, few works directly use sentiment analysis which results to predict the future. People do sentiment analysis, but do not explicitly use the sentiment analysis results to predict the movie success.However in this study we are using sentiment analysis to predict or forecast the box office success. Forecasting is an important aspect that plays a vital role in enabling to take precautionary measure in various fields. It also helps in providing knowledge about the related field that enables in learning what could be the possible results that can be depicted as the outcome of the system that we are analyzing. There have been several methods of forecasting that can be implemented to resolve the issue of unpredictable future outcome. Using this forecast one can act upon the results or on methods to improvise the results in order to ensure the quality of the output generated. One can also be prepared for the fields wherein the results can be applied so as to produce beneficiary products.

Manuscript Received on December 2014.

Sameer Thigale, Department of Computer, MIT College of Engineering, Savitribai Phule Pune University, Pune, India.

Tushar Prasad, Department of Computer, MIT College of Engineering, Savitribai Phule Pune University, Pune, India.

Ustat Kaur Makhija, Department of Computer, MIT College of Engineering, Savitribai Phule Pune University, Pune, India.

Vibha Ravichandran, Department of Computer, MIT College of Engineering, Savitribai Phule Pune University Pune, India.

II. FORECASTING/PREDICTION

Frequently there is a time lag between awareness of an impending event or need and occurrence of that event. This lead time is the main reason for planning and forecasting. If the lead time is zero or very small, there is no need for planning. If the lead time is long, and the outcome of the final event is conditional on identifiable factors, planning can perform an important role. In such situations, forecasting is needed to determine when an event will occur or a need arise, so that appropriate actions can be taken.

In management and administrative situations the need for planning is great because the lead time for decision making ranges from several years (for the case of capital investments) to a few days or hours (for transportation or production schedules) to a few seconds (for telecommunication routing or electrical utility loading). Forecasting is an important aid in effective and efficient planning.

Some of the areas in which forecasting currently plays an important role are:

1. Scheduling: Efficient use of resources requires the scheduling of production, transportation, cash, personnel, and so on. Forecasts of the level of demand for product, material, labor, financing, or service are an essential input to such scheduling.
2. Acquiring resources: The lead time for acquiring raw materials, hiring personnel, or buying machinery and equipment can vary from a few days to several years. Forecasting is required to determine future resource requirements.
3. Determining resource requirements: All organizations must determine what resources they want to have in the long term. Such decisions depend on market opportunities, environmental factors, and the internal development of financial, human, product, and technological resources. These determinations all require good forecasts and managers who can interpret the predictions and make appropriate decisions.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

A wide variety of forecasting methods are available to management. These range from most naive methods, such as use of the most recent observation as a forecast, to highly complex approaches such as neural nets and econometric systems of simultaneous equations.

To deal with the diverse range of applications, several techniques have been developed. These falls into two major categories: quantitative and qualitative methods. These categories can be summarized as:

1. Quantitative: Sufficient quantitative information is available.
   - Time series – Predicting the continuation of historical patterns such as the growth in sales or gross national product.
   - Explanatory – Understanding how explanatory variables such as prices and advertising affect sales.
2. Qualitative: Little or no quantitative information is available, but sufficient qualitative knowledge exists.
   - Predicting the speed of telecommunications around the year 2020.
   - Forecasting how a large increase in oil prices will affect the consumption of oil.
3. Unpredictable: Little or no information is available.
   - Predicting the effects of interplanetary travel.
   - Predicting the discovery of a new, very cheap form of energy that produces no pollution.

III. METHODS OF FORECASTING

Quantitative forecasting can be applied when three conditions exist: 1. Information about the past is available. 2. This information can be quantified in the form of numerical data. 3. It can be assumed that some aspects of the past pattern will continue into the future.

We will be using the quantitative method to forecast the box-office success of the movies as all the related data can be framed in a quantified manner. An additional dimension for classifying quantitative forecasting methods is to consider the underlying model involved. There are two major types of forecasting models: time series and explanatory models.

Explanatory models assume that the variable to be forecasted exhibits an explanatory relationship with one or more independent variables. The purpose of the explanatory model is to discover the form of the relationship and use it to forecast future values of the forecast variable. Explanatory model consists of methods like simple regression and multiple regression, etc.

Time series forecasting treats the system as a black box and makes no attempt to discover the factors affecting its behavior. Therefore prediction of the future is based on past values of variables and/or past errors, but not on explanatory variables which may affect the system. The objective of such time series forecasting is to discover the pattern in the historical data series and extrapolate that pattern into the future.

1. Naïve method – Uses last period's actual value as a forecast.
2. Simple Mean (Average) – Uses an average of all past data as a forecast.
3. Simple moving average – Uses an average of a specified number of the most recent observations, with each observation receiving the same emphasis (weight).
4. Weighted moving average – Uses an average of a specified number of the most recent observations, with each observation receiving a different emphasis (weight).
5. Exponential smoothening – A weighted average procedure with weights declines exponentially as data become older.
6. Trend projection – Technique that uses the least squares method to fit a straight line to the data.
7. Seasonal indexes – A mechanism for adjusting the forecast to accommodate any seasonal patterns inherent in the data.

IV. LITERATURE SURVEY

There are various models available for forecasting. We are here using the regression model for better accuracy and efficiency. There are many other models such as weighted average model, Exponential smoothening, Trend projection which might be a less efficient as in case of expected outcomes.[1][10].

The various factors that could be considered for calculating the success rate might be attention seeking, Distribution, Polarity, Type of film etc. The concept of sentiment analyses can be used for calculating the polarity factor.[2].

Sentiment analysis deals with checking the positivity and negativity of a given sentence thereby determining the emotion that the user is portraying. Using sentiment analysis we can determine the ratio of the positive, negative and neutral tweets.[11][12].

The attention seeking factor and distribution factor can be calculated using the twitter data.[3].

There exists other factor that may contribute for predicting the success rate. Star cast, Sequel, category of the film may be some of those.[5][6]

V. MODEL EXISTING BEFORE

People have previously used regression method of forecasting as a tool for predicting the revenue of a product using social media. This model can also be applied for the success prediction of a given movie. This model can be generalized for various fields including share market. To begin with consider the data collected regarding the product over time, in the form of reviews, user comments and blogs. Collecting the data over time is important as it can measure the rate of chatter effectively. The data can then be used to fit a liner regression model using least squares. The parameters of the model include:

- A : rate of attention seeking
- P : polarity of sentiments and reviews
- D : distribution parameter

Let y denote the revenue to be predicted and ϵ the error. The liner regression model can be expressed as:

$y=ßa*A+ ßp*P+ ßd*D+\epsilon$ (1)

where the ß values correspond to the regression coefficients.

The attention seeking parameter captures the buzz around the product in social media. In this model we can express how the rate of tweets on Twitter can capture attention on movies accurately. This coefficient has been found to be most significant in previously performed experiments. The polarity parameter relates to opinions and views that are disseminated in social media. After the release of the movie this gains importance and adds to the accuracy of the predictions. In the case of movies, the distribution parameter is the number of theatres a particular movie is released in. In the case of other products, it can reflect their availability in the market.

A. Flaws of the model

Although the generalized model can be used for predicting the revenue there are many other aspects of twitter data that can be considered for calculating accurate results in an efficient manner. The movie and various genres of them should also be considered as it can be proved to be a catalyst for calculating the hype created. More over data like the twitter follower count, number of users, no of tweets can be proved to be beneficial. The time of release and its effect might also decide how successful it might be. The holiday effect and the day on which the movie is released play a vital role. All these factors might increase the accuracy of the model. The tweets in various lingual scripts should be considered which is not included in the generalized model.

VI. MODEL PROPOSED

In this study, we are utilizing regression methods for forecasting the box office success. We checked for the flaws in the generalized model and have thereby considered few other factors in order to overcome them. We intend to collect the tweets on hourly basis using Twitter API. We then segregate the data as number of tweets, no of users, no of followers of each tweet. Based on the streaming twitter data we calculate the contributing factors that add up to the success rate. We calculate the contributing factors [1][3] with the help of the data acquired in every hour. At the end of twelve hours using the summation of these factors we calculate the regression coefficients. And at the end of 24 hours we will calculate the outputs considering the coefficients. As mentioned earlier we will have 7 values of the regression output which we will use to obtain the error factor and thereby the corrected and accurate regression outputs. These output values are plotted against time. By analyzing the nature of the graph we conclude the success rate of the movie.
The proposed model can be mathematically represented as follows:

$$Y = ß_aA + ß_pP + ß_dD + ß_bC + ß_eE + ß_sS + \epsilon \quad (2)$$

Where

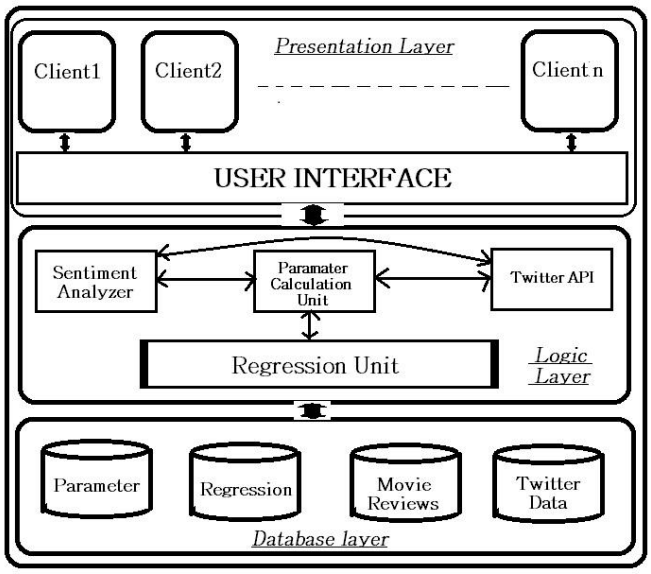
- A : rate of attention seeking
- P : polarity of sentiments and reviews
- D : distribution parameter
- C : Category
- E : Star cast
- S : Sequel
- ϵ : Error Factor

The proposed model consists of the following aspects:

- Three Tier Architecture
- Multiple liner regression model
- Critical Period
- Contributing Factors

A. Three Tier Architecture

The three tier architecture helps in making the system transition more transparent to the user as shown in figure 1. The user here can be the theatre owner or the producer of the movie who intends to maximize the revenue generated from the movie by creating hype amongst the mob. The three tier architecture simplifies the load on the server and increases the interaction with the client. In our model, the presentation layer which is the topmost layer consists of the user requirements. The user creates his account where he is issued with a login Id and Password. He inputs the movie name and the release date of the movie, the occasion on which the movie is released if any. He also enters few other factors such as the genre of the movie, Distribution area etc. Soon after this the user is expected to give commands so as to begin with the process of forecasting. The user can also discard a record if the movie has crossed a certain amount of time beyond the critical period. The Logic layer is where the entire regression process is carried out and the final forecasting is done. The database layer is for recording the data inputs and forecasting results. If forecasting for a given movie already exists in the database layer then the results are directly given to the user in the presentation layer, else the forecasting process is followed by using the logic in logic layer and stored in the database present.



SYSTEM ARCHITECTURE

Figure 1

B. Multiple Liner Regression Model

A multiple linear regression model is a linear model that describes how a y-variable relates to two or more x-variables (or transformations of x-variables).In multiple regression there is one variable to be predicted, but there are two or more explanatory variables.
The general form of regression model is given as:

$$Y_i = ß_0 + ß_1X_{1,i} + \ldots\ldots + ß_kX_{k,i} + \epsilon_i \quad (3)$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Where Y_i , $X_{1,i}$,…, $X_{k,i}$ represent the ith observations of each of the variables Y , X_1, ,…, X_k respectively, $ß_0$,$ß_1$…,$ß_k$ are fixed (but unknown) parameters and ϵ_i is a random variable that is normally distributed with mean zero and having a variance σ_ϵ^2.

There are several assumptions made about X_i and ϵ_i which are important:

1. The explanatory variables X_1, ,…, X_k take values which are assumed to be either fixed numbers (measured without error), or they are random but uncorrelated with the error terms ϵ_i. In either case, the values of X_j(j=1,2,...k) must not be all the same.
2. The error terms ϵ_i are uncorrelated one another.
3. The error terms ϵ_i all have mean zero and variance σ_ϵ^2 , and have a normal distribution.

Now, in practise, the task of regression modelling is to estimate the unknown parameters of the model, namely $ß_0$,$ß_1$…,$ß_k$, and σ_ϵ^2. From a known dataset, the LS procedure can be applied to determine b_0,b_1…,b_k. Thus the pragmatic form of the statistical regression model is as follows:

$$Y_i=b_0+b_1X_{1,i}+\ldots\ldots+b_kX_{k,i}+e_i \quad (4)$$

• The estimates of the β coefficients are the values that minimize the sum of squared errors for the sample.
• The letter b is used to represent a sample estimate of a β coefficient. Thus b_0 is the sample estimate of β_0 , b_1 is the sample estimate of β_1, and so on.
• A predicted value is calculated as $\hat{y}_i = b_0+b_1X_{1,i}+\ldots\ldots+b_kX_{k,i}$, where the b values come from statistical software and the x-values are specified by us.
• A residual (error) term is calculated as

$$e_i = y_i - \hat{y}_I \quad (5)$$

the difference between an actual and a predicted value of y.

C. Critical Period

Forecasting mechanism of our model is time bound. The ideal period for which the regression is constrained to is called the critical period. The accuracy of the model increases by approaching the end of the critical period. The end of the critical period also depicts the end of the regression model used and the accurate results which can be further used for predicting the success rate. The critical period in this model is of 3 weeks. Two weeks before the release date and one week after the release date. The contributing factors are calculated and checked for every week during the critical period. If the movie record has past its' critical period and is old enough then the user might delete the record of the movie thereby making the system memory efficient.

D. Contributing Factors

These are the factors that add up to the success rate of the movie. Forecasting mechanism is carried out on the basis of the contributing factors. Ideally these factors are considered within a specified critical period in order to obtain accurate and efficient results. This time period is an important factor binding the entire mechanism. The contributing factors that we have used in our model are as follows:

- Attention Seeking parameter/Hype
- Polarity
- Distribution factor
- Sequel, star cast and category
- Holiday effect

1. *Attention Seeking Parameter/Hype*

Attention seeking parameter decides the hype created amongst the mob regarding the release of the upcoming movie. It is an extremely important factor that contributes the success of the movie. With appropriate promotion of the movie the number of audience might increase. The release of the music album or the poster of the movie usually creates the hype amongst the mob. This increases the excitement for the release of the movie and thereby increasing the posts or the tweets regarding the movie on the social networking sites. Social Media plays a vital role in increasing and calculating the hype factor and to analyze it so as to obtain accuracy in the same. Here are major concern is the tweets and comments shared on the largely popular social networking site Twitter. As mentioned earlier we access the tweets regarding the movie which is about to be released. The various promotional activities act as a catalyst in increasing the number of tweets for the particular movie. We need to sort and arrange the data from Twitter API for the analysis of the same. The factors considered are as follows:

- Number of tweets from all users
- Number of distinct users
- Follower count of a particular user
- Rate of the tweets

Using the above mentioned parameters the attention seeking parameter is calculated after every hour. The twitter data has to be segregated and stored in the form of table in the database layer of the architecture. Then by using the formula mentioned below [4] we calculate the value.

$$\frac{\text{No of distinct users}}{\text{no of tweets by all users}} + \text{Rate of Tweets} \quad (6)$$

By adding up the contributing factors we come up with the value q [2] which will be used in calculation of the regression coefficient for the hype factor at the end of every twelve hours and update the output. This coefficient can provide an appropriate result to know the attention gathered by the movie amongst the people. The accuracy of the parameter increases with the end of critical period.

2. *Polarity*

Polarity is a factor that determines the sentiments involved in the tweet of the user. Sentiment Analysis is a branch wherein the sentiments of the user are analyzed from the written scripts or recorded voice. If positive sentiments are reflected in the sentence then the sentiment analyzer marks the sentence as positive. Similarly in the case of evidence of negative statements in marks it as negative. There may also be a scenario wherein the sentence might neither reflect positive nor negative emotions. Such sentences are termed as neutral by the analyzer. The polarity determines the ratio of the positive, negative and neutral tweets. It reflects the emotions that the user wishes to convey through the tweet regarding the given movie. The hype created results in number of tweets which in turn allows us to segregate and break the tweets into positive and negative tokens.

There are various tools available in order to carry out sentiment analysis such ad Rapid Miner, Lingpipe, Sentiword etc.In this study we use Sentiword for breaking the sentence into tokens and label them to be positive, negative or neutral. Then using the formula mentioned [5][6] we can calculate the polarity after every hour and thereby the factor q[2]. After a period of twelve hours we calculate the regression coefficient for the same with the help of q. We also reduce the possibility of ambiguity by adding acronyms and symbols used in slang to the dictionary by looking at which the sentiment analyzer decides whether the sentence reflects positive or negative emotion, so that the tweets in slang language can also be analyzed and can be used.

Polarity = Positive sentiment tweets/Negative sentiment Tweets (7)

In case of neutral tweets we consider:

Subjectivity = [Positive &Negative tweets]/[Neutral tweets] (8)

3. Distribution Factor

Distribution factor depends on the area wherein the movie is being released over a specified period of time. It can also be considered as the reach amongst the mob. It is one of the factors affecting the q [2] value. The factors it considers for are follower count and average no of followers per all users who tweeted. The distribution factor can also be calculated as follows:

Distribution Factor (D) = follower count-t/follower count (9)

Where t =average no of followers per all users who tweeted
Using the above formula and q [2] value we calculate the regression coefficient for this factor after every twelve hours.

4. Sequel, Star cast and Category

In this factor the various aspects of the movie to be released are categorized and considered. In case of the Sequel we make use of dummy variable where if it is a sequel it is assigned a value '1' else '0'.Similary for category there exists for category of movies namely comedy, horror, action, romance, animation. We use four dummy variable setting the values as '1' in case of that category else '0'. In case of the star cast, we check for the number of followers of the star cast and accordingly sort them as popular, debutant and not popular. Two Dummy variables are used in this case that work in similar fashion as mentioned above. The regression coefficients are calculated for the above factors after every 12 hours with the help of least squares method and q [2] value.

5. *Holiday Effect*

The effect of holidays or festive seasons on the box office success of a movie can be easily handled using variable holiday effects (indicator variables). As we know festivals like Diwali and Dasehra always occur around the month of October/November or Christmas always occurs in December. So we use seasonal dummy variables to include the effect of Diwali in the October/November component.
We have dummy variables for each of the months as D1, D2… D11. The value for each variable can be defined as:

D1 = 1, if there occurs a festival in the month of January else it is 0,
D2 = 1, if there occurs a festival in the month of February else I is 0,
...and so on.

To avoid the problem of multicollinearity, we use one less than the number of variables I.e. 12-1 = 11 dummy variables. The co-efficient associated with these variables reflect the average difference in the forecast variable between these months and the omitted month. All the above mentioned factors are substituted in equation [3] so as to obtain the final result of the regression model. Then these results are plotted against time in order to analyze the success rate

VII. ALGORITHM PROPOSED

Algorithm:

1. Start
2. Accept command for movie scheduling from user.
3. If already scheduled
   - **3.1** Send report
   - **3.2** Goto step 5
4. Generate required database for the movie
   - **4.1** Collect tweets from twitter API
     - **4.1.1** Segregate positive and negative tweets
     - **4.1.2** Calculate PN ratio i.e polarity for movie
     - **4.1.3** Calculate tweets/hour ratio
   - **4.2** Calculate other factors D,B
   - **4.3** Update the existing database
5. Check time elapsed
   - **5.1** After 12 hours
     - 5.1.1 Goto step 4.1
   - **5.2** After 24 hours
     - **5.2.1** Calculate regression co-efficients
     - **5.2.2** Calculate the final output "y"
     - **5.2.3** Plot graph of x v/s y.
   - **5.3** After 1 week
     - **5.3.1** Send report
   - **5.4** After 3 weeks
     - **5.4.1** Stop regression
     - **5.4.2** Send report
6. If user sends a delete command
   - 6.1 Flush all data related to the movie
7. Stop

VIII. CONCLUSION

The model proposed in this study will produce better results than the existing models related to this problem area. In this project we have shown how social media can be utilized to forecast future outcomes. Specifically, using the rate of chatter from tweets from the popular site Twitter, we constructed a multiple linear regression model for predicting box-office revenues of movies in advance of their release. At a deeper level, this work shows how social media expresses a collective knowledge which, when properly tapped, can yield an extremely powerful and accurate indicator of future outcomes. The accuracy of our model increases with end of critical period approaching.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

REFERENCES

1. FORECASTING-Methods and Applications by- Spyros M., Steven W., Rob H., Edition(3).Wiley Publication.
2. Sitaram Asur&Bernardo A. Huberman,"Predicting the Future with Social Media", Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, pp. 492-499,Oct 2010.
3. A.Reddy,P,Kasat,A.Jain, St.Francis Institute of Technology,"Box-Office opening prediction of Movies based on Hype Analysis through Data Mining", International Journal of Computer Application(0975-8887),Volume 56-No 1,October 2012
4. Minxue Huang and Feng Wang,"Using Online WOM to Forecast Box Office for Movies Coming Soon",Wireless communcations, networking And Mobile Computing,2008. WiCOM'08.Fourth International Conference. Oct 2008.
5. Seonghoon Moon, Suman Bae, Songkuk Kim ,"Predicting the Near-Weekend Ticket Sales Using Web-Based External Factors and Box-Office Data",Web Intelligence(WI) and Intelligent Agent Technologies(IAT), 2014 IEEE/WIC/ACM International Joint Conferences,Aug 2014
6. Andrei Oghina, Mathias Breuss, Manos Tsagkias&Maarten de Rijke,"Predicting IMDB movie ratings using social media", Proceedingsof the 34th European conference on Advances in Information Retrieval, pp. 503 507,November 2013
7. Jure Leskovec, Lada A.Adamic and Bernado A. Huberman,"The dynamics of viral marketing" In proceedings of the 7th ACM Conference on Electronic Commerce,2006.
8. Lyric Doshi,"Using Sentiment and Social Network Analyses to predict Opening-Movie Box Office Success",Department of Electrical and computer MIT,USA,Feb 2010.
9. David Jensen and Jennifer Neville,"Data Mining in Social Networks",Computer Science Department,University of Massachusetts,Amherst.
10. Swart,William,"Demand Forecasting With Multiple Rgression", Developed exclusively for IEEE eLearning Library,Dec 2011
11. Neethu,Rajsree,R.,"Sentiment Analysis In Twitter Using Machine Learning Techniques",Computing,Communicationsand Networking Technologies(ICCCNT),2013 Fourth International Conference,Jully 2013.
12. Singh,V.K,Priyani,R.Uddin,A,Waila,P,"Sentiment Analysis of Movie Reviews:A New Feature-Based Heuristic for Aspect-Level Sentiment Classification",Automation,Computing,Communication,Control and Compressed Sensing,(iMac4s),2013 International Multi-Conference,March 2013.

AUTHORS PROFILE

Sameer Thigale, Student Bachelor of Engineering, Department of Computer,MIT College of Engineering,Savitribai Phule Pune University, Pune, India.

Tushar Prasad, Student Bachelor of Engineering Department of Computer,MIT College of Engineering,Savitribai Phule Pune University, Pune, India.

Ustat Kaur Makhija, Student Bachelor of Engineering Department of Computer,MIT College of Engineering,Savitribai Phule Pune University, Pune, India.

Vibha Ravichandran, Student Bachelor of Engineering Department of Computer,MIT College of Engineering,Savitribai Phule Pune University, Pune, India.