Comment Volume Prediction using Regression

Mandeep Kaur
Computer Science engg.CTIEMT
Jalandhar,Punjab,India

Prince Verma
Computer Science engg.CTIEMT
Jalandhar,Punjab,India

ABSTRACT

The latest decade lead to a unconstrained advancement of the importance of online networking. Due to the gigantic measures of records appearing in web organizing, there is a colossal necessity for the programmed examination of such records. Online networking customer's comments expect a basic part in building or changing the one's acknowledgments concerning some specific indicate or making it standard. This paper demonstrates a preliminary work to exhibit the sufficiency of machine learning prescient calculations on the remarks of most well known long range informal communication site, Facebook. We showed the customer remark patters, over the posts on Facebook Pages and expected that what number of remarks a post is depended upon to get in next H hrs. To automate the technique, we developed an item display containing the crawler, information processor and data disclosure module. For prediction, we used the Linear Regression model (Simple Linear model, Linear relapse model and Pace relapse model) and Non-Linear Regression model(Decision tree, MLP) on different data set varieties and evaluated them under the appraisal estimations Hits@10, AUC@10, Processing Time and Mean Absolute Error.

Keywords

Social media, Comment volume, Pace regression, REP Tree.

1. INTRODUCTION

Initially Internet was explored for limited purposes as for reading, watching or shopping only. But today's consumers are more intelligent and are utilizing platforms such as blogs, content sharing sites, wikipedia and social networking for multiple- purposes as for news, advertisements, communication, commenting, banking, marketing etc. Social media user's comments are also very much influential in order to make or change people's perception and to bring some topic in trending which can also affect a firm's reputation, survival and sale's revenues.

Facebook is getting the position of being most popular social networking site, as in Facebook, the daily data ingested into the databases, is beyond 500 terabytes. Figure 1. shows the popularity of different networks across the world. These are ranked by the number of active accounts by January 2016.

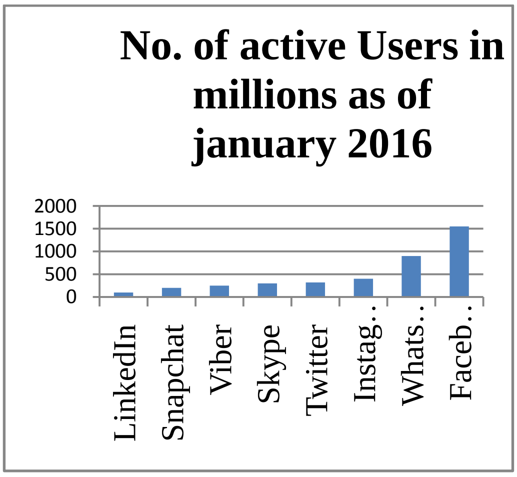

Fig. 1. Number of active users in millions and social networks

Active users are those which have logged in during the last 30 days. With a constant presence in the lives of their users, social networks have a strong social impact [19].

With such a huge popularity, facebook is having a big amount of data which is impossible to analyze manually and so, there is a huge need for the automatic analysis of such data. For this analysis, we built up a software model comprising of the Crawler, Data processor and Information Analytic module. The Analytic component is oriented towards the feedback volume prediction that a post is expected to receive by considering the properties of domain.

This paper is organized as Section II discuss about the Social media Comments and its significance and Section III discuss about the related works, Domain Specific Concepts are discussed in Section IV, Problem formulation is discussed in Section V. Section VI and Section VII discusses about the experimental settings and Results. The paper is closed with Conclusion and Future work in Section VII.

2. SOCIAL MEDIA COMMENTS

Comments are the opinions of social media users about a document. These are usually short textual messages referring to the main text of the document or to other feedbacks.

The nature and valence of user's comments on various social media platforms plays a significant role in building or changing the one's perceptions regarding some specific topic or to making it popular. Even, in today's era, it has an important place in various fields whether it is education, sales or predictions etc.

i. Perceptions--The paper[15] examined that the perception of people is influenced by the comments of users on social media. The findings were explored in the context of relationship status, in which a person made an announcement about the formation or breakup of a real-world relationships on Facebook. The results suggested that the comments from different users affects the perceptions of a Facebook relationship status update. It showed that the positive comments lead to favourable attitudes whereas negative comments lead to poorer attitudes towards the status. And also the observer's attitude towards a relationship status are more driven by the count of the comments than the real nature of the status, as positive facebook status can be seen as negative if the associated comments are negative in nature.

ii. Education- This work[16] researched the impacts of utilizing an online networking web application to help learning and educating in classrooms on learners execution. By looking at the proportion of number of Facebook posts and remarks for instructive purposes to the one for non-instructive purposes, a conclusion is drawn that the utilization of Facebook has an effect on understudy's learning exhibitions, as the understudies who invested more energy in instruction related posting and remarking earned better evaluations

iii. Predicting the future-This paper[17] demonstrate that the contents of social media can be used for the predictions of real-world outcomes. Predictions were performed on platform Twitter.com, using almost 3 million tweets. It predicted the box-office revenues of movies by constructing a linear regression model, in advance of their release dates. The results outperformed in terms of accuracy and concluded that a topic's rank in future is strongly correlated with the amount of attention it receives on social media.

iv. Driving force in s/w evolution- User feedback is important in improving software Quality. Many software companies collect data on user satisfaction through various means. The user's feedback available on the sites thereby supply software companies with a rich source of information that can be used to improve future releases. Even Apple's App Store stated that the newest release addresses many of the issues raised by users. Therefore, user feedback can be and has been the driving force in software evolution [13].

v. Trending- It shows a list of topics that have recently spiked in popularity which is evaluated based on the volume of the response towards a particular topic/issue or product.

vi. Sales and Marketing- McDonald's Australia is serving more than 1.7 million individuals crosswise over more than 940 Australian eateries every day. This worldwide fast food chain has utilized Facebook video promotions to share its remarkable Australian story, interfacing with more than 5 million individuals in only 5 weeks [18]. And even Coca-Cola Korea has used age group targeting and video ads to connect with 4.5 million young consumers, in order to promote a new pineapple beverage, Sunny 10, which increased its purchase intent for the product by 2 points.

3. RELATED WORKS

The related works to our research are:

The paper [2] has developed an industrial proof-of-concept demonstrating the automatic analysis of the documents on Hungarian Blogs. In this paper, author has trained various regressor models by considering various features of the blogs and measured the results by using the parameters Hits@10 and AUC@10. The result shows that the regression models outperform than naive models.

Even, classifiers can be used to categorize the comment volumes in specific classes like paper [3].It reports on predicting the comment volume of news articles before their publication using Random Forest Classifier based on the set of five features i.e. surface, cumulative, textual, semantic, and real-world features .It addresses the task in two steps- first binary classification of articles with the potential to receive comments and second to classify articles with 'low volume' and 'high volume'. Outcomes show better results for binary classification and evaluated that the Textual and semantic features are strong performers among others. In the similar way, paper [7] has made analysis on the content and publish time of online news agencies, to detect effective factors of diffusing contents in public. It has also used the Random Forest Classifier to classify articles in three categories i.e. without commented, moderately commented(1-6) and highly commented(>6).The proposed model has made predictions with more than 70% accuracy and reports that the publish date and a weight introduced for content measure, were most informative features. The results can be refined by considering important days (i.e. elections, festivals, holidays) and geographical features in prediction. While paper [5] has shown the dynamics of user generated comments on seven different news websites, using the log-normal and the negative binomial distributions and predicted the comment volume using Linear model and enable comparison across various news sites. The results showed that prediction of long term comment volume is possible with small error after 10 source-hours observations.

The paper [4] has worked on Social bookmarking website Digg.com. By using comment information, it defined a co-participation network between users and studied the behavioral characteristics of users. It measured the entropy and inferred that the users at Digg are interested in wide range of topics. Using a classification and regression framework, it has predicted the popularity of online content based on comment data and social network derived features. It reported a one to four percent loss in classification accuracy while predicting the popularity metric by using only first few hours of comment data as compare to all the available comment data. The results can further be improved by analyzing the polarity of the comments.

Various topic models can be used to extract the hidden topics in post's content. The paper [6] has worked on political blogs using Latent Variable topic model and analyzed the relationship between the content and comment volume. It has also used Naïve Bayes model for binary prediction task i.e. high volume or low volume and evaluated the prediction using precision, recall and F1 measures. It concluded that the modeling topics can improve recall while predicting high volume posts. Even paper [8] has predicted the formation of user-to-content links in Flickr Groups to predict the chance that a user will comment or like an image updated by another user. It has taken into account both the community effect using Transactional Mixed Membership Stochastic Block (TMMSB) model and content effect using Latent Dirichlet Allocation(LDA) for predicting user-to-content links. The time zone effects can be used in future in order to make results more accurate.

Summarization of the user's comments is even more difficult task as these are usually mixed with different opinions, specifically in case of restaurants where different opinions refer to different dishes but evaluated as an overall score of restaurants. Paper [10] has presented a new approach for comment summarization in the context of restaurants. It used the real-world comments, crawled from Yelp and Dianping, the most popular English and Chinese restaurant review web sites. Using the attributes of the dishes and the user's remarks on the attributes as two independent dimensions in the latent space, it constructed a bilateral topic model which is combined with the opinionated word extraction and clustering-based selection algorithms, it provides a high-quality summary on the restaurants as well as the dishes served by the restaurants. This concept can be further used for wider applications like for various selling goods or services.

In contrast to these works, we have focused on leading platform i.e. Facebook, a leading platform and targeted the regression models as Pace Regression and REP Tree.

4. DOMAIN SPECIFIC CONCEPTS

For this work, the concepts related to our domain are: (1) Source : source refers to the page that produces the post. (2) Links : these are the pointers to other related posts or pages refered in main text or comments. (3) Main text : the text refers to the main topic of the post. (4) Comments : these are the opinions of the users about a post or other comments mentioned under the main text. (5) Feedback Volume : volume of feedback can be measured as the count of words in the comment section, the number of comments, the number of distinct users who leave comments, or a variety of other ways. These measures can be affected by various factors like main text of post, link to other posts, the time of day the post appears, a side conversation, Page likes, page check-ins, page talking about or page category etc. (6) Feedback Volume Prediction : the user comment patterns are modeled over the posts/documents appeared in the past and based on it, predictions are made on the number of comments that the posts/documents are expected to receive in next coming hours. (7) Base-time : This is used to simulate the scenario to make predictions of the post after the selected time i.e. base time, it is simulated in the sense, as the real values of feedback are already know which will be used further to evaluate the predictions of the models. (8) Variants : For effective time-lined analysis, the variants are used. Variant-X defines that X variants are there in the dataset for particular instance. Weight for particular instance is also increased as we are considering the same instance X-times by varying its base selected base date/time.

5. PROBLEM FORMULATION

For our work, we refer the predictive modeling techniques and address this work like a regression problem. For Predicting the feedback, the patterns of user's comments are modeled over the posts appeared in the past and based on it, a model is trained and predictions are made on the expected number of comments that a post may receive in next N hours. Figure 2. depicts the work flow of our process.

A. *Data Pre-processing :*

1) Crawler : Pages are crawled from facebook in the raw form using language JAVA.

2) Data Cleansing : Data is cleaned by discarding the data which has missing values, the data which is older than three days with respect to the selected base time or the posts with no comments.

3) Data Splitting : Data splitting is done on the temporal basis to divide it into training and testing data. In Training data, the value of the target is already known and used to train the model, and based on this information, we want to predict the value of the target for those cases where it is unknown using which feedbacks of testing data is predicted.

4) Vectorization : Data cannot be used in the raw form directly. For analysis, it is needed to be transformed in vector forms. In this data is transformed from document form(.txt) to the vectors form(.xls/.arff) which can be easily processed by weka tool.

B. Features set :

Analysis is performed using the various properties of the application into account. We considered following features as input to the predictor and take 1 feature as output value for each post. The various features are:

1) Page Features: We recognized 4 elements of this class incorporates highlights that characterize the ubiquity/Likes, classification, checkin's and discussing of wellspring of page. Page likes : It is an element that characterizes clients support for particular remarks, pictures or pages. Page Checkin's : It is a demonstration of indicating nearness at specific spot. Page Category : It characterize the class of the page as Entertainment, sports, educational, music band, movies etc. Page Talking About : This is the real check of clients who are "locked in" and cooperating with that Page. The clients who really return to the page, after liking the page once. This incorporate exercises, for example, remarks, likes to a post, shares by guests to the page.

2) Post Features: This incorporate some report related elements like length of archive, time hole between chosen base date/time and record distributed date/time. It ranges from (0; 71), report advancement(promotion) status values (0,1) and share count of post. 5 elements of this class are recognized.

3) Comment related Features: This incorporates the example of remark on the post in different time interims with respect to the randomly chose base time, named as C1 to C5. C1: Total comment number before chose base date/time. C2: Comment number in last 24 hrs w.r.t to chose base date/time. C3: Comment check is last 48 hrs to last 24 hrs w.r.t to base date/time. C4: Comment check in initial 24 hrs after the distribution of the report/post, yet before the chose base date/time. C5: The distinction in the middle of C2 and C3. Besides, we aggregate these components by source and built up some determined elements by computing min, max, normal, middle and Standard deviation of 5 aforementioned highlights. Along these lines, including the 5 key elements and 25 inferred fundamental elements, we got 30 components of this class.

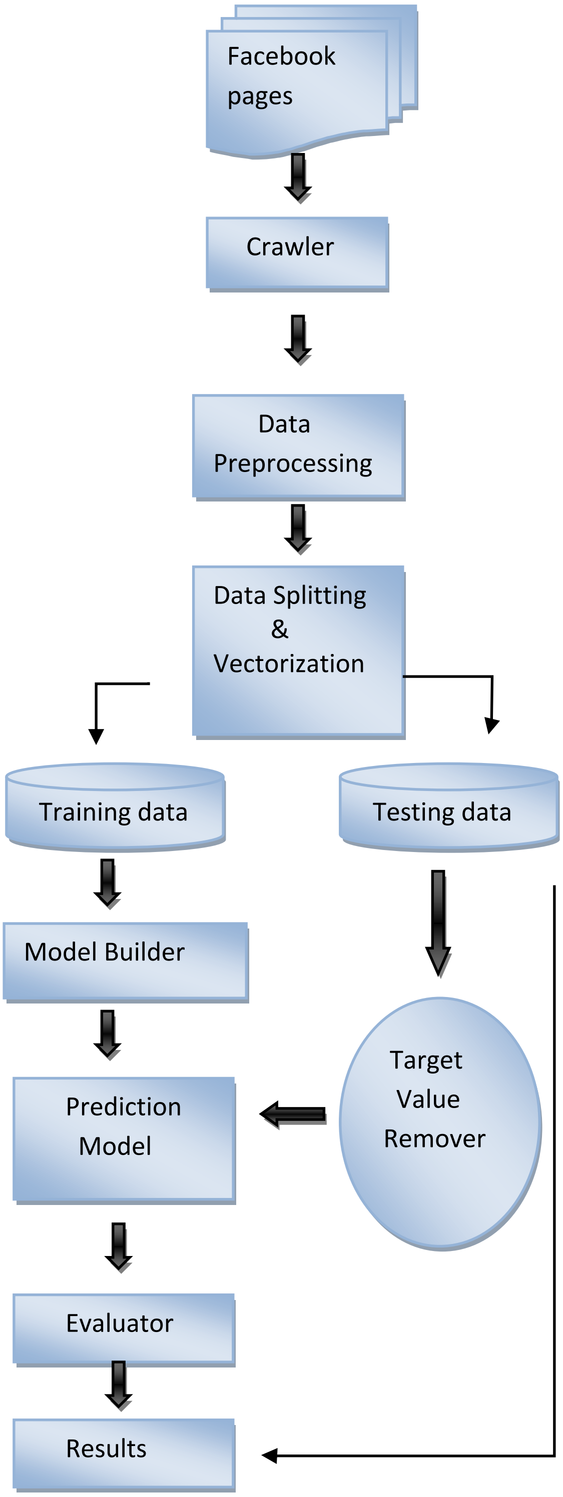

Fig 2. Flow of Feedback prediction process.

4) Timeline Features: Binary pointers (0,1) refers to the day of the week on which the post was distributed and the day on which the prediction is to compute. 14 elements of this write are distinguished.

C. Machine Learning for Feedback Prediction

In order to predict the feedback, we selected some time in the past and reproduce as though the present time would be the chosen time. We call the chose time as baseTime. As we realize what happened after the baseTime, i.e., we know what number of feedbacks the post got in the following H hours after baseTime, we know the estimations of the objective for these cases. At the same time, we just consider pages that were distributed in the most recent 3 days in respect to the baseTime, as more older pages typically don't get any new feedback.

This prediction problem is addressed by Regression Analysis. In our prototype we used Linear Regression models i.e. Simple Linear Regression, Linear Regression, PACE regression and Non-Linear Regression Models i.e. REP Tree, Multi-Layer Preceptron model.

Regression Analysis is a statistical modeling technique for finding the relationship between the target variable(Y) and the predictor variables(X).

$$Y \approx f(\mathbf{X}, \boldsymbol{\beta})$$

β is the coefficient of X and it defines the rate of change of target variable (Y) with one unit of variation in predictor variable(X).Variables used to explain the target variable are called explanatory variables or independent variables and the variables which are explained, are called response variable or dependent variable. It is mostly used for Forecasting and Prediction. In our case, the explanatory variables are the features of facebook pages and the response variable is the feedback volume.

Regression Models are classified as:

1) *Simple Regression Model :* If there is one predictor variable only to define the response variable, then the model is called a Simple Regression Model.

$$Y = b_o + b_1X_1 \qquad (1)$$

2) *Multiple Regression Model :* It is a model when there are more than one predictor variables available to define the target variable, then the model is called a Multiple Regression Model.

$$Y = b_o + b_1X_1 + b_2X_2 + ... + b_kX_k \quad (2)$$

Regression analysis is carried out either by Linear Regression or Non linear Regression techniques.

1) *Linear regression:* In Linear regression, the regression function is defined by the linear combination of finite number of unknown parameters (β), which are estimated from given data.

2) *Nonlinear regression:* It is a regression analysis technique, in which data is modeled by a regression function which is a combination of nonlinear parameters. While the equation of a linear model has a fundamental structure, nonlinear mathematical statements can take a wide range of structures as it uses logarithmic, trigonometric and exponential functions, which is the reason nonlinear regression gives the most adaptable curve fitting usefulness. The data are fitted by a method of successive approximations.

Nonlinear regression is similar to linear regression as both seek to track the response from a given set of variables, in a graphical form. While nonlinear models are more complicated to develop because the data are fitted through a method of series of approximations.

Pace Regression is a linear regression method that enhances established Ordinary Least Squares (OLS) regression by assessing the impact of every variable and utilizing a clustering analysis to enhance the statistical basis for evaluating their contribution to the whole regression. Additionally, it perform very well than other linear modelling methods and subset selection methods, which look for a diminishment in dimensionality that drops out as a natural characteristic of pace regression[21].

Decision Tree is a non-linear method which is used for both regression and classification. Decision tree encodes the rules in the form of if- then-else which predict the target value based on the given features. This formation of rules is called Decision tree learning. The importance of the decision tree modeling lies in the fact that Decision trees can model the regression and classification function of any type. But it is prone to the problem of overfitting which can be easily handled by Reduced Error pruning tree. It is a simple and speedy procedure for pruning decision trees, given by Quinlan. In this, internal nodes are transverse from leaves to the root and at each node it checks whether replacing the node with most frequent class effects its accuracy or not. The node is pruned if it does not affect the accuracy. This procedures is repeated on each node in the upward direction until it decreases the accuracy[22].

Artificial Neural Network is a powerful data driven, self-adaptive, flexible computational tool which has the capability of capturing non-linear and complex characteristics with high degree of accuracy.

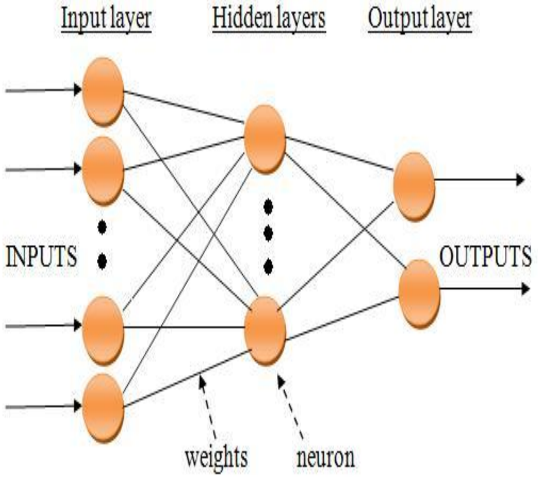

Fig. 3 Block Diagram of Multi-layer Preceptron

Multilayer perceptron (MLP) is a classifier/regressor depending on the feed-forward artificial neural network. MLP involves numerous layers of nodes. Each and every layer is completely linked to the upcoming layer in the network. Nodes in the input layer represent the input data. All other nodes maps inputs to the outputs by performing linear composition of the inputs with the node’s weights w and bias b and applying activation function. Multilayer perceptron (MLP) is an information processing technique based on biological nervous systems process information, such as in brain

6. EXPERIMENTAL SETTINGS

For our analysis, we crawled the data from facebook pages for preparing the training and testing set of our proposed model. In all out two thousand five hundred pages are crawled for 60,000 posts and it contains 4,100,500 comments utilizing Facebook Query Language (FQL). The crept information includes upto certain Giga bytes. The crept information is cleaned and we cleared out with 50,000 posts. We split the cleaned corpus into two subsets utilizing temporal split as 80% for training data and 20% for testing data and then these datasets are sent to next modules for further processing.

For Training the data, PACE Regression and Decision Tree (REP Tree) models are used through WEKA(The Waikato Environment for Knowledge Analysis). On the other hand for Testing the data : Out of the testing set, 10 test cases are created randomly with 100 occurrences in each for assessment and afterward they are changed to vectors.

A. Evaluation Parameters:

The performance of the models is evaluated using following parameters :

1) Hits@10: It is an accuracy parameter for the proposed work. In this, we consider the main 10 posts having largest number of remarks/comments according to the results of our prediction. And then, we compare and count the number of posts among these that had received highest number of comments in real. After that, it is averaged over all the test cases [2].

2) AUC@10: It tells about the precision of the predictions. AUC means area under the receiver-operator curve. For this, we consider the 10 posts receiving the largest number of comments in actual, as positive. Then, these posts are sorted according to the predicted count of comments and AUC is calculated. It is written as:

$$AUC = \frac{TP}{TP + FP}$$

where TP: True positives, FP: False positives.

3) M.A.E: Mean Absolute Error is an average of the absolute errors. It is used to measure how close forecasts or predictions are to the actual outcomes of comment volume.

The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|.$$

where f_i is the prediction and y_i the true value.

4) Evaluation time: It is the time taken by the model to perform the evaluation.

7. RESULTS AND DISCUSSION

The performance of the evaluated models is depicted in Table 1and 2, using measures as Hits@10, AUC@10, Evaluation Time & M.A.E on implemented models.

Results using Linear regression Models:

A. Hits@10 :-

For this measure, graph shows that the Pace regression performs best among the linear regression models with 6.4+-01.114hits(in case of variant-4), where Linear regression performs better than Simple linear regression with maximum response of 6.1+-00.943hits (in case of variant-4). On the other hand, Simple linear regression performs lowest with 4.4+-01.200hits in each case of variants.

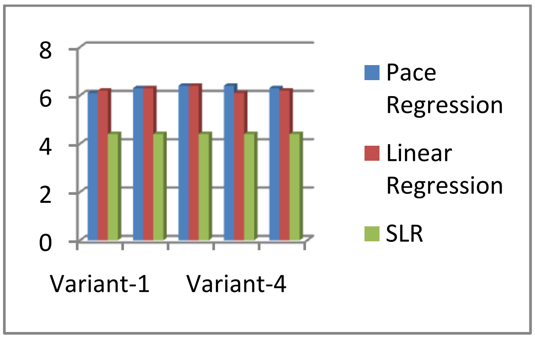

Fig 4. Hits@10

B. AUC@10 :-

For AUC@10 metric results shown in Figure5 demonstrate that Pace Regression has maximum accuracy with value 00.888+-00.022 in variant-1 and Linear regression performs better with 00.887+-00.024 value in variant-1 than Simple linear regression 00.681+-00.029 (in variant-1).

Table 1 Experimental Results Using Linear Models

MODELS	PARAMETERS	Variant – 1	Variant – 2	Variant -3	Variant – 4	Variant – 5
Pace Regression	Hits@10	6.1+-01.136	6.3+-01.100	6.4+-01.020	6.4+-01.114	6.3+-01.005
	AUC@10	00.888+-00.022	00.884+-00.023	00.869+-00.032	00.879+-00.023	00.881+-00.024
	Time Taken	35.589Sec	55.879Sec	80.655Sec	112.320Sec	132.883Sec
	M.A.E	20.850%	15.549%	04.307%	17.088%	19.908%
Linear Regression	Hits@10	6.2+-01.077	6.3+-01.005	6.4+-01.020	6.1+-00.943	6.2+-01.249
	AUC@10	00.887+-00.024	00.881+-00.027	00.866+-00.039	00.877+-00.028	00.878+-00.029
	Time Taken	34.627Sec	73.087Sec	91.006Sec	112.276Sec	138.922Sec
	M.A.E	20.516%	11.644%	01.722%	18.420%	21.231%
Simple Linear R.	Hits@10	4.4+-01.200	4.4+-01.200	4.4+-01.200	4.4+-01.200	4.4+-01.200
	Auc@10	00.681+-00.029	00.683+-00.027	00.680+-00.029	00.677+-00.032	00.674+-00.035
	Time Taken	08.724Sec	17.034Sec	25.437Sec	33.500Sec	41.729Sec
	M.A.E	25.813%	27.555%	25.956%	24.668%	23.204%

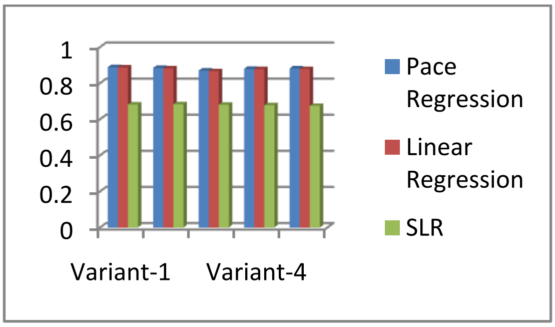

Fig 5. AUC@10

C. M.A.E :-

From the graph shown in Figure6, Linear Regression is gaining the minimal error 01.722% (variant-3), where pace regression produces atleast 04.307%(variant-3) of error which is better than 25.956% in case of Simple linear regression in variant-3.

D. Evaluation Time :-

From Figure 7, it is clear that Simple linear regression gives the best performance in terms of execution time with maximum value of 41.729Sec for Variant-5 followed by Pace regression with value 132.883Sec under variant-5, where Linear regression had taken 138.922Sec for evaluation of variant-1.

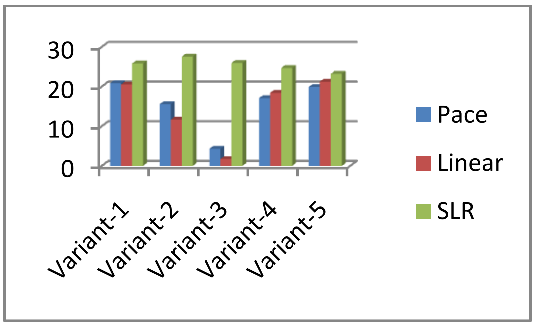

Fig 6. M.A.E.

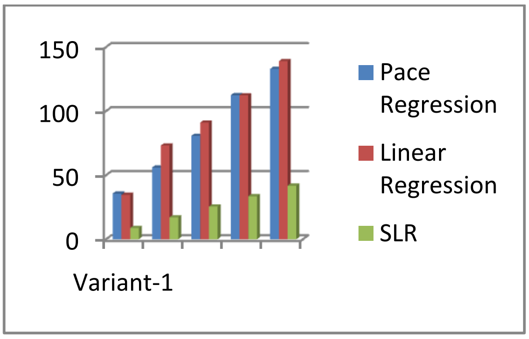

Fig 7. Evaluation Time.

Table 2 Experimental Results Using Non-Linear Models

MODELS	Parameters	Variant – 1	Variant – 2	Variant -3	Variant – 4	Variant – 5
REP Tree	Hits@10	6.3+-00.781	6.4+-00.917	6.8+-00.872	6.1+-01.375	6.6+-01.281
	AUC@10	00.797+-00.093	00.692+-00.136	00.670+-00.095	00.700+-00.098	00.606+-00.084
	Time Taken	32.819Sec	80.487Sec	118.613Sec	179.430Sec	222.646Sec
	M.A.E	09.799%	22.576%	32.331%	28.203%	42.699%
MLP(20, 10)	Hits@10	5.9+-01.136	6.1+-01.375	6.2+-01.249	6.0+-01.612	6.2+-00.980
	AUC@10	00.856+-00.060	00.726+-00.104	00.628+-00.150	00.745+-00.126	00.758+-00.131
	Time Taken	1346.406Sec	2221.599Sec	3639.18Sec	5988.576Sec	5525.13 Sec
	M.A.E	31.843%	30.837%	44.803%	27.425%	27.063%

Results using Non-Linear regression Models:

A. Hits@10 :-

For Hits@10, the REP tree have shown maximum response of 6.8+-00.872 in variant-3 which is far better than that of MLP's maximum response i.e.: 6.2+-01.249 in variant-3. And graph shows that the REP tree performs better than the MLP network.

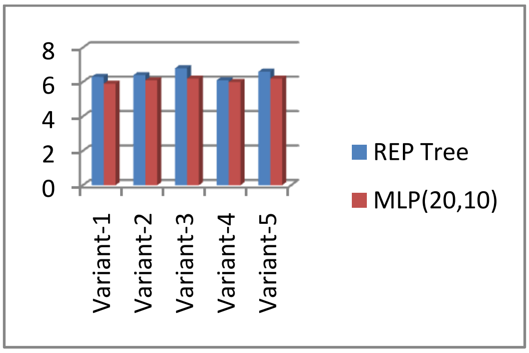

Fig 8. Hits@10

B. AUC@10 :-

For AUC@10 graph is shown in Figure 9, the performance of REP Tree and MLP is almost equivalent with nominal differences in the values with highest scores i.e. 00.856+-00.060 in variant-1(in case of MLP) and 00.797+-00.093 in variant1(in case of REP Tree).

C. M.A.E :-

From the graph shown in Figure 10, M.A.E is minimum is in case of REP Tree that is 09.799% for variant-1, where MLP has 27.425 of mean absolute error in variant-4.

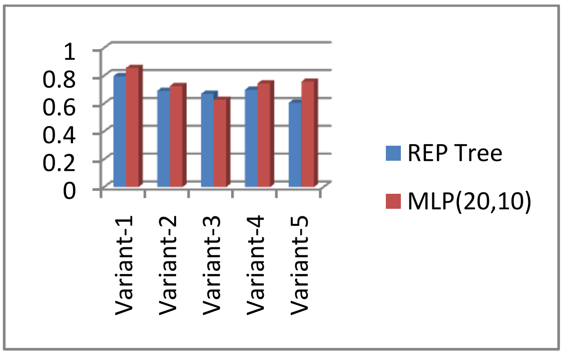

Fig 9. AUC@10

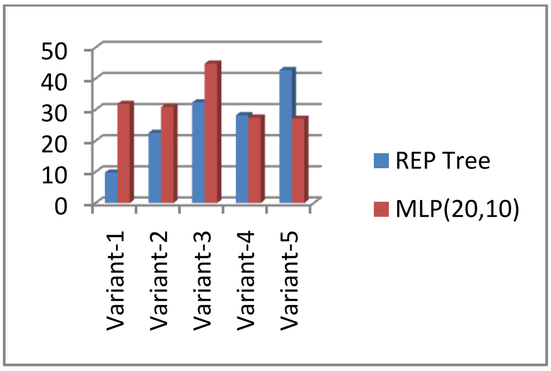

Fig 10. M.A.E.

D. Evaluation Time :-

By analyzing the results, it is observed that the REP Tree with variant-1 have given result in 32.819Sec which is far better than MLP with time 1346.406Sec for variant-1.

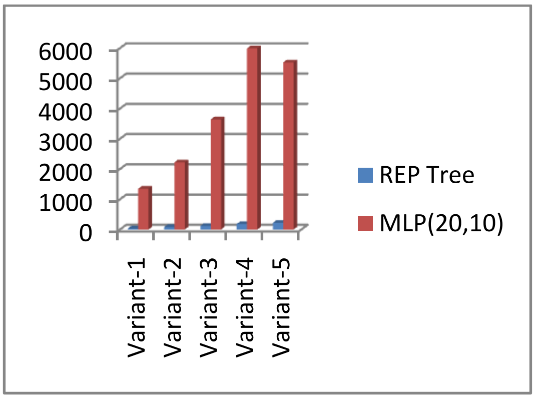

Fig 11. Evaluation Time

8. CONCLUSION AND FUTURE SCOPE

Comment volume prediction being a measure to analyze user's interest in a Facebook post is latest and trending area of research. By examining various linear and non-linear predictive models, we came to the conclusion that in the linear models, PACE regression performed better than others as it filters the features for prediction using OLS and thresholding. In Non-linear models REP tree performed better than the other non-linear models due to its divide and conquer strategy, minimal prediction time and pruning to reduce overfitting. It is also observed that the variants have no significant effect over prediction.

9. REFERENCES

[1] S. M. Tan, P.N, V. Kumar, Introduction to data mining, Pearson Addison Wesley Boston, 2006.

[2] Buza Krisztian, "Feedback Prediction for Blogs", Springer International Publishing on Data Analysis, Machine Learning and Knowledge Discovery,2014, pp. 145-152. doi:10.1007/978-3-319-01595-8 16.

[3] M.Tsagkias, W. Weerkamp, M. de Rijke, "Predicting the Volume of Comments on Online News Stories",CIKM'09 Proceedings of the 18th ACM conference on Information and knowledge management , pp.1765-1768, 2009.

[4] Jamali, S. and Rangwala, H., "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis", Web Information Systems and Mining, IEEE International Conference,2009, pp. 32-38. doi: 10.1109/WISM.2009.15.

[5] M. Tsagkias, W. Weerkamp, M. de Rijke, "News Comments: Exploring, Modeling, and Online Prediction", ECIR'2010 Proceedings of the 32nd European conference on Advances in Information Retrieval, Springer, pp.191-203,2010.

[6] Yano, Tae, and Noah A. Smith. "What's Worthy ofComment? Content and Comment Volume in Political Blogs" In 4th International AAAI Conference on Weblogs and Social Media, 2010.

[7] Balali, A. and Rajabi, A. and Ghassemi, S. andAsadpour, M. and Faili, H., "Content diffusion prediction in social networks", Information and Knowledge Technology (IKT), 5th IEEE Conference,2013, pp. 467-471. doi: 10.1109/IKT.2013.6620114.

[8] Negi, S. and Chaudhury, S., "Predicting User-to-content Links in Flickr Groups", Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference, 2012 pp. 124-131. doi: 10.1109/ASONAM.2012.31.

[9] Rahman, M.M. , "Intellectual knowledge extraction from online social data", Informatics, Electronics Vision (ICIEV), IEEE International Conference, 2012, pp. 205-210. doi:10.1109/ICIEV.2012.6317392

[10] Rong Zhang, Zhenjie Zhang, Xiaofeng He, Aoying Zhou, "Dish Comment Summarization Based on Bilateral Topic Analysis" Data Engineering (ICDE), 31st IEEE International Conference, 2015, pp. 483 – 494. doi: 10.1109/ICDE.2015.7113308

[11] S. M. Tan, P.N, V. Kumar, Introduction to data mining, Pearson Addison Wesley Boston, 2006.

[12] Z.C. Khan , Thulani Mashiane, "An analysis of Facebook's Graph Search", Information Security for South Africa(ISSA), IEEE, pp.1-8, 2014.

[13] Laura V. Galvis Carreno, Kristina Winbladh "Analysis of User Comments: An Approach for Software Requirements Evolution",35th International Conference on software Engineering ,USA, IEEE, 2013, pp. 582-591. doi: 10.1109/ICSE.2013.6606604.

[14] Yu-Hsiu, Hsueh-Yi Lai "Effects of Facebook Like and Conflicting Aggregate Rating and Customer Comment on Purchase Intentions", Springer International Publishing in Universal Access in Human-Computer Interaction. Access to Today's Technologies, 2015,pp.193-200.doi: 10.1007/978-3-319-20678-3_19.

[15] Paul W. Ballantine , Yongjia Lin, Ekant Veer, "The influence of user comments on perceptions of Facebook relationship status updates", Computers in Human Behavior, Elsevier 2015, Volume- 49, pp.50–55. doi:10.1016/j.chb.2015.02.055

[16] N. Leelathakul, K. Chaipah, "Quantitative Effects of using Facebook as a Learning Tool on Students' Performance" 10th International Joint Conference on Computer Science and Software Engineering (JCSSE),IEEE 2013, pp.87 – 92. doi: 10.1109/JCSSE.2013.6567325.

[17] Sitaram Asur, Bernardo A. Huberman, "Predicting the Future With Social Media" Web Intelligence and

Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference, pp. 492 – 499. doi: 10.1109/WI-IAT.2010.63

[18] https://www.facebook.com/business/success

[19] http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[20] Jan H. Kietzmann , K.Hermkens, Ian P. McCarthy, B. S. Silvestre, “Social media? Get serious! Understanding he functional building blocks of social media” Business Horizons, Elsevier 2011, Volume- 54, Issue 3, pp. 241-251. doi:10.1016/j.bushor.2011.01.005S

[21] Wang Y, Witten IH (1999) Pace regression. Technical Report 99/12, Department of Computer Science, The University of Waikato

[22] Tapio Elomaa, Matti Kaariainen, “An Analysis of Reduced Error Pruning”, Department Journal of Artificial Intelligence Research 15 (2001) 163-187.