International Journal of Database Theory and Application
Vol.9, No.11 (2016), pp.71-82
http://dx.doi.org/10.14257/ijdta.2016.9.11.07

Fusion of PACE Regression and Decision Tree for Comment Volume Prediction

Mandeep Kaur[1] and Prince Verma[2]

Dept. of Computer Science engg.CTIEMT
Jalandhar, Punjab, India
[1]shergill2mk@gmail.com,[2]princeverma.cse@gmail.com

Abstract

The analysis of social networking sites is a vast area of research as there are tremendous measures of records showing up in online networking. Predicting the comment patterns of users on these sites is a complex decision making process. This paper proposes a hybrid model of linear regression (PACE regression) and non linear regression (REP Tree) that predicts the likelihood of the comment volume, which a post may receive by analyzing the various features of the corresponding page, post and previous records of comment patterns of users. To mechanize the procedure, a model is built comprising of the crawler, data processor and information revelation module. The new hybridized model has improved the time and space complexity along with Accuracy by building a right sized tree using only significant features with low misclassification rate.

Keywords: *Social media, Comment volume prediction, PACE regression, REP tree*

1. Introduction

Social networking sites have turned into a prominent stage for individuals to share information and communicate with others, which can essentially influence a company's reputation, survival and deal's incomes. These services are going about as a multi-device with routine applications, *e.g.* news, commercials, communication, remarking, banking, advertising and so on.

Social media Comments are the opinions of users in the form of short textual messages which refers to the main topic of the post. In today's era, the semantic and the count of user's feedback plays an influential role to build or change the perceptions of users [1]. So, analyzing the comment patterns of social sites is an important area for research.

Being the most active social networking site [2], the 'Facebook' has been used in order to estimate the volume of comments that a post is expected to receive in the next few hours. With such a huge popularity, Facebook is having a big amount of data which is impossible to analyze manually and so, an automated system is required for analysis of such data. And for analysis, a software model is built up comprising of the Crawler, Data processor and Information Analytic module. The Analytic component is oriented towards the prediction of comment volume that a post is expected to receive by considering the properties of the domain. The various concepts that are of interest are:

- *Source*: it refers to the page that produces the post.
- *Links*: these are the pointers to other related posts or pages referred in the main text or comments.
- *Main text*: the text refers to the main topic of the post.
- *Comments*: these are the opinions of the users about a post or other comments mentioned under the main text.
- *Comment Volume*: Volume of feedback can be measured as the count of words in the comment segment, the quantity of comments, the number of distinct

users/clients who leave remarks, or a number of other ways. These measures can be affected by various factors like main text of the 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 and so forth.

- *Comment Volume Prediction*: the user comment patterns are demonstrated over the posts/documents showed up in the past and in view of it, expectations are made on the quantity of remarks that the posts/archives are relied upon to get in next coming hours.
- *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 known which will be used further to evaluate the predictions of the models.
- *Variant*: For effective time-lined examination, the variations are used [3]. Variant-X defines that X variants are there in the dataset for the 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.

2. Related Work

In the related attempts to the examination, paper by K. Singh, *et al.* [3] has dealt with Facebook, utilizing decision trees and neural networks and built up a product model exhibiting the Comment volume forecast. It has made assessments utilizing different dataset variations and reasoned that the Decision trees performed well than the Neural Networks. And Buza K. has built up a mechanical verification of-idea, in paper [4] exhibiting the programmed examination of the records on Hungarian Blogs. In the paper, different regressor models are modeled by considering different elements of the web journals and assessed the outcomes utilizing Hits@10 and AUC@10 measures. The outcome demonstrates that the regression models beat than naive models. Out of the different types of features, the basic features (including aggregated features by source) seem to be the most predictive ones. Out of the distinctive features, the basic elements appear to be the most prescient ones.

Indeed, classifiers can be utilized to classify the comment volumes in particular classes as in paper [5] by M. Tsagkias *et al.* It provides details regarding foreseeing the remark volume of news articles by utilizing Random Forest Classifier in view of the arrangement of five components features *i.e.* textual, surface, semantic, cumulative and real-world features. It addresses the work in two stages- first the characterization of articles with the possibility to get remarks and second to order articles with 'low volume' and 'high volume'. Results indicate the better outcomes for binary classification and assessed that the textual and semantic elements are solid features among others.

In the comparable way, paper by Balali A. and Rajabi A. *et al.* [6] has made examination on the publication time and content of online news offices, to identify effective variables of diffusing content out. It additionally utilized the Random Forest Classifier to characterize articles in three classifications, *i.e.* without remarking, modestly remarked (1-6) and profoundly commented (>6).The proposed model has made expectations with more than 70% exactness and reports that the distribute date and a weight presented for the measure of content, were most instructive components. The outcomes can be refined by considering essential days (*i.e.* elections, celebrations, occasions) and geological elements in prediction.

While paper [7] by M. Tsagkias, W. Weerkamp and M. de Rijke has demonstrated the progression of client created remarks, utilizing the negative binomial distributions and the log-normal. It also anticipated the remark volume, utilizing the Linear model and empower examination crosswise over distinctive seven news sites. The outcomes demonstrated that forecast of long term remark volume is feasible with little error by observing 10 sources.

In paper [8], Jamali, S. and Rangwala, H. has made analysis on Social bookmarking site Digg.com. By utilizing the comment data, it characterized a co-support system amongst clients and contemplated the behavioral qualities of clients. It quantified the entropy and gathered that the clients at Digg are keen on an extensive variety of themes. Using the regression and classification systems, it has anticipated the fame of online data is based on remark information and on features of a social group. It reported a 1 to 4% loss in the accuracy of classification while foreseeing the fame metric by observing the comment information for just initial couple of hours as contrast with all the accessible comment information.

Yano, Tae, and Noah A. Smith. has made analysis on political blogs in paper [9] and analyzed the relationship between the comment volume and content using Latent Variable topic model. It has also made a binary prediction using the Naïve Bayes model for classifying volume into two classes, *i.e*. high volume or low volume and measure the performance of prediction under precision, recall and F1 parameters. It demonstrates that modeling the topics can enhance recall in case of high volume posts.

Most well known communication site, Facebook is utilized by Rahman, M. M. in paper [10] for analysis and used the data mining engineering to gather social information. This paper has gathered distinctive traits, for example, about me, remarks, wall post, age from Facebook utilizing Facebook API key and mined the information to learn the correlation on age bunch premise for different uses as job responsibility distribution, human conduct expectation, pattern recognition, product promoting and decision making. It has utilized the K-nearest neighbor calculation to classify the value attributes such as age count, wall count, interest count and music count to classify them into various class levels.

User’s comments summarization is a more difficult task as these are generally mixed with various opinions, particularly in the case of restaurants where diverse opinions refer to various dishes but assessed as a general score of restaurants. Paper [11] by Rong Zhang *et al.* has introduced another methodology for summarizing the comments of restaurants. It utilized the present reality remarks gathered from most prominent English and Chinese eatery survey sites *i.e*. Yelp and Dianping. The author has mapped the properties of the dishes and the client's comments on these properties onto the two dimensions independently in the space. Then, developed a topic model which is a combination of opinionated word extractor and clustering algorithm which provides a good quality of review’s summary on the eateries and the dishes served by the eateries. This idea can be further utilized for more extensive applications like for marketing and services.

3. Comment Volume Prediction

In this work, predictive modeling technique is referred and addressed like regression problem.

3.1. Problem Formulation

For Predicting the volume of comments, 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, which is 24 hours in our case. Figure 1 depicts the work flow of our process.

3.1.1. Facebook Data

A Crawler is programmed for fetching the raw data from Facebook pages. This collected raw data is then preprocessed for discarding the data which is older than three days with respect to the selected base time or the posts with no comments.

3.1.2. Data Splitter

Data is split into the training (80%) and testing part (20%), on the basis of time. After splitting, Data is transformed in vector form *i.e.* in .arff format to make it suitable for analysis by weka tool.

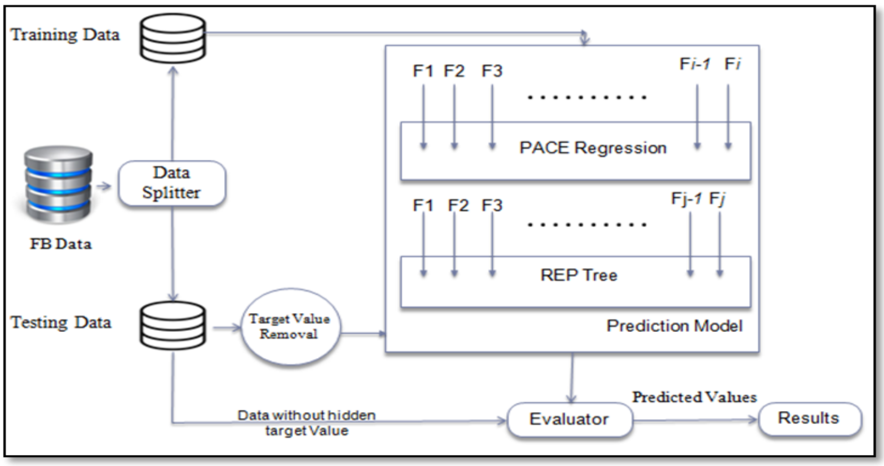


Figure 1. Flow of Comment Volume Prediction Process

- Along with Vectorization, Variants are generated in which a Base time is selected to simulate the scenario for predictions. Data before the Base time is used to train the model and based on this data feedback is predicted for next hours. But the data after Base time represents the actual values of the feedback, which are used to evaluate the performance of the model.
- Variant - X, defines that, X instances are derived from single training instance, as described in example of Facebook official page id: 177344652298186 with post id: 816365251729453. It received a total of 373 comments at time of crawling as shown in Figure 3.

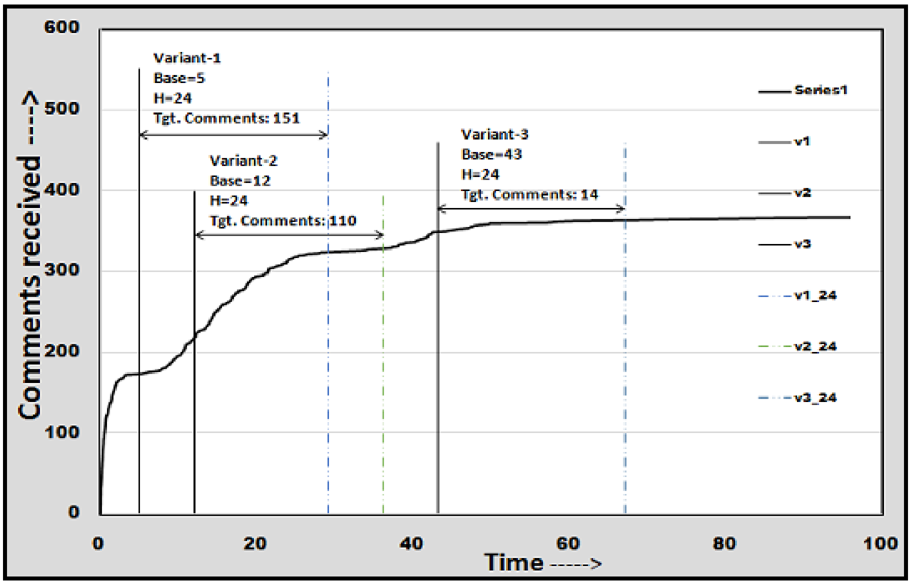


Figure 2. Flow of Volume of Comments with Time

- The Projection Adjustment by Contribution Estimation (PACE) regression is a Linear Regression strategy. It is used for modeling the relationship between a target variable (Y) and one or more input variables (x).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{1}$$

β is the coefficient of x and it models the relationship by defining the impact on target variable (Y) with one unit of variation in the associated input variable (x). β is calculated using the formula:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2}$$

It gives the least square solution to the problem and anticipate the unknown estimation of a target variable from the known estimation of the input variable. In our case x1, x2….,xk refers to the various features discussed in table I and Y is the volume of comments.

Thus, it seeks a selection procedure that falls out as a natural byproduct of the PACE regression [12]. It evaluates the significance of each feature with respect to the target variable and discards the insignificant features by zeroing the corresponding coefficients (β). Being the linear and parametric model, PACE catches all the information inside its parameters and foreseen the future based on these parameters.

- Rep Tree [13] is further used for feedback predictions using only filtered features by PACE.

It is a decision tree which is used both for regression and classification. It builds the model in the form of tree structure by breaking down the dataset into smaller homogeneous subsets (*i.e.* instances with similar values). In classification, it calculates the highest Information gain as the splitting criteria of an attribute as:

$$\mathrm{Gain}(T,X) = \mathrm{Entropy}(T) - \mathrm{Entropy}(T,X) \tag{3}$$

It is based on the decrease in entropy after a dataset is splitted on an attribute. The sample with zero entropy is completely homogeneous. Entropy is calculated as:

$$E(T,X) = \sum_{c \in X} P(c)\, E(c) \tag{4}$$

and

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{5}$$

While in Regression, Standard Deviation Reduction is used in place of Information gain as in case of classification.

$$\mathrm{SDR}(T,X) = \mathrm{SD}(T)\text{-}\mathrm{SD}(T,X) \tag{6}$$

where SD is the standard deviation of the attributes and evaluated as:

$$SD = \sqrt{\frac{\sum(x-\mu)^2}{n}} \tag{7}$$

The algorithm is run recursively on the non-leaf branches, until all data is classified. The leaf nodes represent the final results by calculating the average of the values if more than one instance are left at leaf nodes.

- After building a decision tree using training dataset, in Reduced error pruning tree, post pruning is performed using validation dataset (*i.e.* pruning set). Validation set is a subset of training dataset whose target results are already known, it is used to check the accuracy of the tree. The number of correct classifications and misclassifications are recorded at each node. Then, traversing the tree from the leaves to the root and at each node

- Calculate the error which is the sum of all the errors of its child nodes.
- Calculate it again with the same example, after replacing it with leaf node.
- Prune the node with the highest error reduction.
- Repeat it until there is no reduction in error.

Thus pruning helps to reduce the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting. And it makes the tree simple and speedy.

The proposed technique for predicting the Comment Volume is implemented as follows:-

Step1:Pass every feature to PACE Model.
Step2:Find the significance of all the features using the coefficient values (β).
Step3:Filter out the insignificant features with β=0.
Step4:REP Tree receives the significant features which is passed by PACE model.
Step5:REP Tree is constructed on the basis of Standard Deviation Reduction.
Step6:Pruning is performed on the basis of Reduced Error to reduce the overfitting.
Step7:Leaf Nodes contain the Comment Volume outcomes.

3.1.3. Features

For the first issue, *i.e.*, for transforming the records into vectors, we find the following features from each document:

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) Parent features: The first four features defined in the table refers to the parent features as it defines the likes, category, check-ins *etc.* features which belongs to the source page of corresponding post. It shows the role of the page's popularity in bringing the response to the post published on that page.
2) Comment related features: This incorporates five to thirty four features mentioned in the table. Collaboration of Comment Count (CC) features and their derived features (5-29) extracts the pattern of response that a post has received at different times by defining how the comment volume increased or decreased relative to the base time of a post. Comment patterns are captured in Comment Count (CC) features which is diagrammatically shown as:

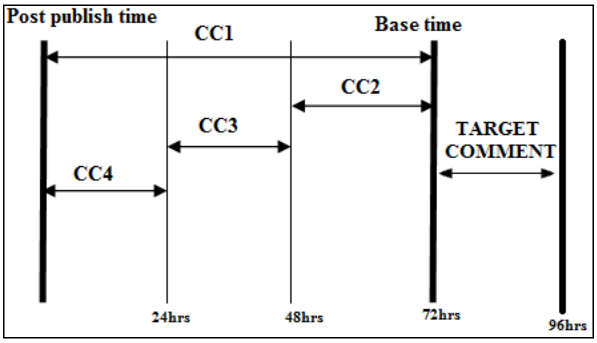


Figure 3. Comment Count (CC) Features

3) Post features: Features (35-39 in the table) define the post related properties like length of archive, *i.e.* how the number of characters or the length of the post brings the attention of users, time hole between chosen base date/time and post published date/time. It ranges from (0-71), report advancement (promotion) status values (0,1) in order to reach to more peoples and share count of the post.
4) Weekday features: In this Binary pointer (0,1) describes the day on which post is published (features 40-46) and the day (features 47-53) for which the prediction is to be calculated.

In the given Table 1, 53 features are identified as Input features and 1 as target value for each post.

Table 1. Features Used

S.no	Features	Description
1	Page Likes	It is the number of individuals following this page and defines the popularity of the page.
2	Page Checkins	It describes how many individuals so far visited this place. This feature is only associated with the places.
3	Page Talking about	It defines the daily interest of individuals towards the posts. This include activities such as comments, likes to a post, shares etc by the visitors of the page.
4	Page Category	It defines the category of the page *e.g.* education, sports, entertainment.
5-29	Derived	These are aggregated features and defined by calculating the min, max, average, median and standard deviation of comment features.
30	CC1	It is the total number of comments before selected base time.
31	CC2	It is the number of comments in last 24 hrs relative to the base time.
32	CC3	It is the number of comments in last 48 hrs to last 24 hrs relative to the base time
33	CC4	It is the number of comments in first 24 hrs after the publication of post but before base time
34	CC5	It is the difference between CC2 and CC3.
35	Base_time	It is the selected time in order to simulate the scenario.
36	Post_Length	It is the number of characters in the post.
37	Post_Share_Count	It is the number of shares of the post.
38	Post_Promotion_Status	To reach more people with posts in News Feed, individual promote their post and this feature act as a binary indictor whether the post is promoted (1) or not (0).
39	H_Local	It describes the number of hrs, for which we evaluate the target comment volume.
40-46	Post_Published_weekday	It represents the day (Sunday….Saturday) on which the post was published.
47-53	Base_DateTime_weekday	It represents the day (Sunday….Saturday) on selected base time.
54	Target_comment	It is the volume of feedback in 24 hrs. after the chose base time.

4. Comment Volume Prediction

For the analysis, the data from Facebook pages is crawled for preparing the training and testing set of the proposed model. In all out two thousand five hundred pages are crawled for 60,000 posts utilizing Facebook Query Language (FQL) and JAVA. Dell, Inc.-Inspiron1545 is used for this evaluation, which is configured with Windows 7 Professional N Operating System, Intel Core 2 Duo CPU with 2.10GHz clock rate processors, with 3.00 GB of RAM and 320 GB of Hard Drive. Evaluation time, may vary with varying system configuration. The crept information includes upto certain Giga bytes. The crept information is cleaned and we cleared out with 50,000 posts. The cleaned corpus is split into two subsets utilizing temporal split as 80% for training data and 20% for testing data and then these datasets are sent to the next modules for further processing.

For training 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 data: 10 test cases are created randomly with 100 occurrences in each for assessment and afterward they are changed to vectors.

4.1. Evaluation Metrics

The performance of the models is evaluated using following parameters:

4.1.1. Hits@10

It is an accuracy parameter for the proposed work. In this, we consider the main 10 posts having the largest number of remarks/comments according to the results of our prediction. And then, we compare and count the number of posts among those that had received the highest number of comments in real. After that, it is averaged over all the test cases [4].

4.1.2. AUC@10

It tells about the precision of the predictions. AUC means an 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 the AUC is calculated [4]. It is written as:

$$AUC = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{8}$$

where TP: True positives, FP: False positives.

4.1.3. M.A.E

Mean Absolute Error is an average absolute of the errors. It is used to measure how close the predictions are to the actual outcomes of comment volume [14]. The mean absolute error is given by:

$$M.A.E. = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{9}$$

where f_i is the prediction and y_i the true value.

4.1.5. Evaluation Time

It is the time taken by the model to perform the evaluation [14].

4.2. Results

The hybrid model is evaluated with three variants of the dataset and results are depicted in Table 2 using measures as Hits@10, AUC@10, Evaluation Time & M.A.E.

4.2.1. Hits@10

For Hits@10, graph shows that the Hybrid Model performs better than PACE model and REP tree model with 7.1 ± 00.833 hits (in case of variant-2) and minimum response of 6.5 ± 00.806 hits (in case of variant-3) which is better than 6.4 ± 01.114 hits value of PACE and 6.4 ± 00.917 value of REP Tree. This shows that the proposed model is performing better in terms of higher accuracy, as shown in Graph:

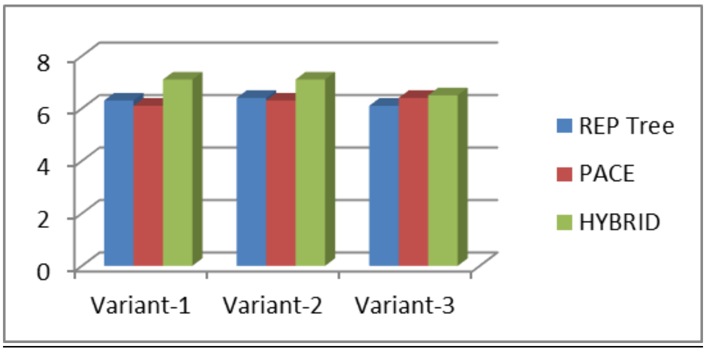


Figure 4. Hits@10

4.2.2. AUC@10

For AUC@10, the results are presented in Figure 5, it demonstrates that the hybrid model is performing better than the base models REP Tree and PACE regression with better precision. The Hybrid model has greater precision with value 00.978 ± 00.024 in variant-1 and PACE regression falls after with 00.888 ± 00.022 value in variant-1 and REP Tree with 00.797 ± 00.093 (in variant-1). Thus, it has been seen that the proposed model is performing better than base model in terms of AUC measure.

4.2.3. M.A.E

From the graph Figure 6, it is depicted that the Hybrid model outperforms well in case of mean absolute error. The Hybrid model has a minimal error with minimal deviation than the REP Tree and PACE regression. Hybrid model is outperforming with minimal error of 12.093 ± 12.060% (variant-3), where PACE regression have error of 20.850 ± 07.360% (variant-1) and REP Tree have 28.203 ± 18.649% (variant-3). Another observation is that for variant-1 the REP Tree have shown good performance but when data is increased *i.e.*: variant-2 and variant-3, the REP's error is increased.

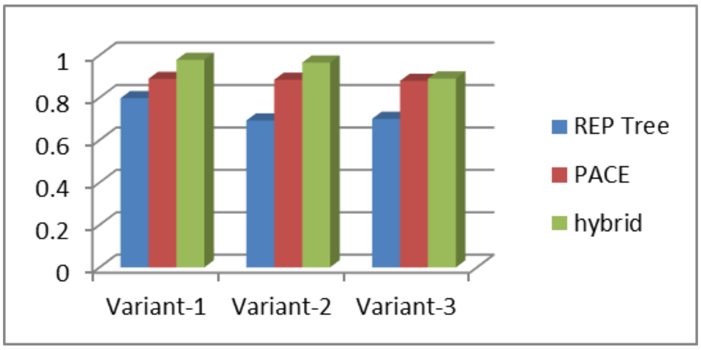


Figure 5. AUC@10

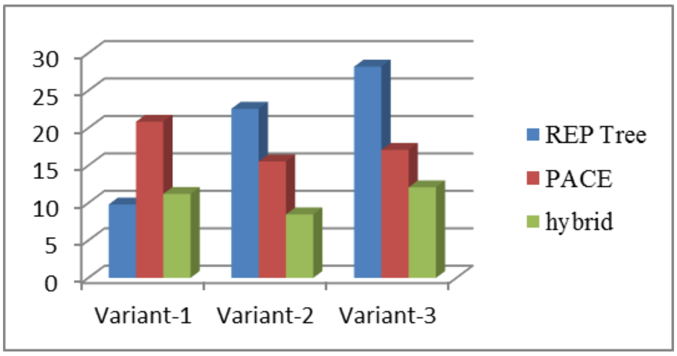
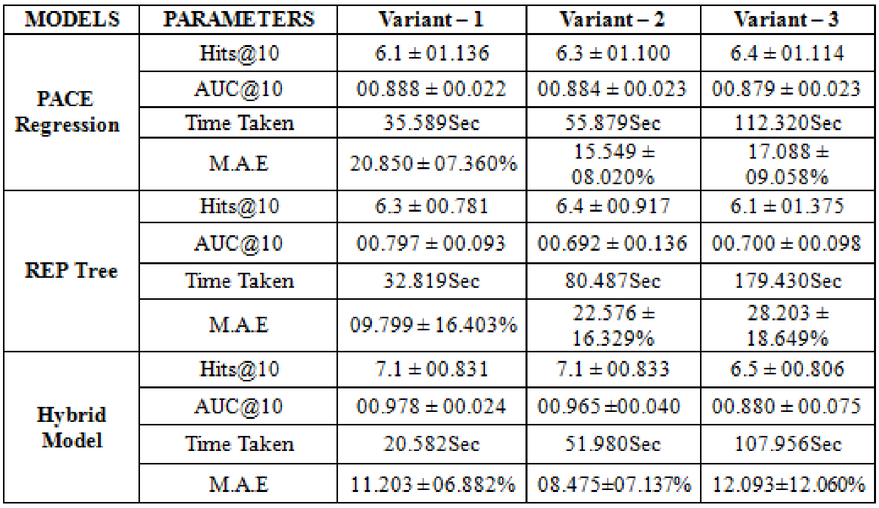


Figure 6. M.A.E

Table 2. Experimental Results



MODELS	PARAMETERS	Variant – 1	Variant – 2	Variant – 3
PACE Regression	Hits@10	6.1 ± 01.136	6.3 ± 01.100	6.4 ± 01.114
	AUC@10	00.888 ± 00.022	00.884 ± 00.023	00.879 ± 00.023
	Time Taken	35.589Sec	55.879Sec	112.320Sec
	M.A.E	20.850 ± 07.360%	15.549 ± 08.020%	17.088 ± 09.058%
REP Tree	Hits@10	6.3 ± 00.781	6.4 ± 00.917	6.1 ± 01.375
	AUC@10	00.797 ± 00.093	00.692 ± 00.136	00.700 ± 00.098
	Time Taken	32.819Sec	80.487Sec	179.430Sec
	M.A.E	09.799 ± 16.403%	22.576 ± 16.329%	28.203 ± 18.649%
Hybrid Model	Hits@10	7.1 ± 00.831	7.1 ± 00.833	6.5 ± 00.806
	AUC@10	00.978 ± 00.024	00.965 ±00.040	00.880 ± 00.075
	Time Taken	20.582Sec	51.980Sec	107.956Sec
	M.A.E	11.203 ±06.882%	08.475±07.137%	12.093±12.060%

4.2.4. Evaluation Time

From Figure 7, it is depicted that the Hybrid model has better performance among other models. For variant-1, the performance of REP Tree is much similar to that of PACE but Hybrid model is performing better. For variant-2 and variant-3, the performance of REP Tree degrades more, than the PACE and Hybrid, whereas hybrid model consistently performed well.

As the hybrid model is an integration of 2 models *i.e.*: REP Tree and PACE regressor, it is performing better as due the hybrid model is result of abstraction of best features of REP and PACE.

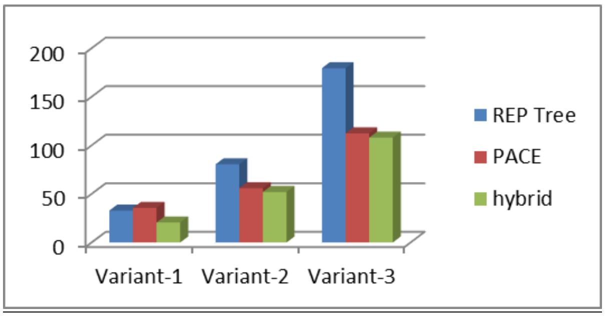


Figure 7. Evaluation Time

4. Conclusion

In this paper, REP tree and PACE regression are fused to predict the comment volume prediction on a Facebook page's post. The PACE is used as a filter to filter out significant features and REP is a prediction model. The performance of the prediction has been estimated and is compared with the existing techniques. The results show that the proposed model yields better predictions. The work can further be extended by including the methods of sentiment analysis in order to find the polarity along with the count of comments expressed under the given post, *i.e*. how many positive/negative or neutral comments that a post may receive in next H hrs.

References

[1] P. W. Ballantine, Y. Lin and E. Veer, "The influence of user comments on perceptions of Facebook relationship status updates", Computers in Human Behavior, Elsevier, doi:10.1016/j.chb.2015.02.055, vol. 49, **(2015)**, pp. 50-55.

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

[3] K. Singh, R. K. Sandhu and D. Kumar, "Comment volume prediction using neural networks and decision trees", in IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015), Cambridge, United Kingdom, **(2015)**.

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

[5] M. Tsagkias, W. Weerkamp and 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, **(2009)**, pp. 1765-1768.

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

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

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

[9] Y. Tae and N. A. Smith, "What's Worthy of Comment? Content and Comment Volume in Political Blogs", In 4th International AAAI Conference on Weblogs and Social Media, **(2010)**.

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

[11] R. Zhang, Z. Zhang, X. He and A. Zhou, "Dish Comment Summarization Based on Bilateral Topic Analysis", Data Engineering (ICDE), 31st IEEE International Conference, doi: 10.1109/ICDE.2015.7113308, **(2015)**. pp. 483-494.

[12] Y. Wang and I. H. Witten, "Pace regression. Technical Report 99/12", Department of Computer Science, the University of Waikato, **(1999)**.

[13] T. Elomaa and M. Kaariainen, "An Analysis of Reduced Error Pruning", Department Journal of Artificial Intelligence Research, vol. 15, **(2001)**, pp. 163-187.

[14] K. Singh, "Facebook Comment Volume Prediction", International Journal of Simulation: Systems, Science & Technology, doi: 10.5013/IJSSST.a.16.05.16, pp. 16.1-16.9.

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

[16] S. M. Tan and P. N. V. Kumar, "Introduction to data mining", Pearson Addison Wesley Boston, **(2006)**.

[17] Z. C. Khan and T. Mashiane, "An analysis of Facebook's Graph Search", Information Security for South Africa (ISSA), IEEE, **(2014)**, pp. 1-8.

[18] L. V. G. Carreno and K. Winbladh, "Analysis of User Comments: An Approach for Software Requirements Evolution", 35th International Conference on software Engineering, USA, IEEE, doi: 10.1109/ICSE.2013.6606604, **(2013)**, pp. 582-591.

[19] Y. Hsiu and H. Y. 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, doi: 10.1007/978-3-319-20678-3_19, **(2015)**, pp.193-200.

[20] P. W. Ballantine, Y. Lin and E. Veer, "The influence of user comments on perceptions of Facebook relationship status updates", Computers in Human Behavior, Elsevier, doi:10.1016/j.chb.2015.02.055, vol. 49, **(2015)**, pp.50-55.

[21] N. Leelathakul and 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, doi: 10.1109/JCSSE.2013.6567325, **(2013)**, pp. 87-92.

[22] S. Asur and B. A. Huberman, "Predicting the Future with Social Media", Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conference, doi: 10.1109/WI-IAT.2010.63, (2010), pp. 492-499.

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

[24] J. H. Kietzmann, K. Hermkens, I. P. McCarthy and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media", Business Horizons, Elsevier, doi:10.1016/j.bushor.2011.01.005, vol. 54, no. 3, **(2011)**, pp. 241-251.

[25] M. Kaur and P. Verma, "Review on Comment Volume Prediction", IOSR Journal of Computer Engineering (IOSR-JCE), doi: 10.9790/0661-180203134138, vol. 18, no. 2, **(2016)**, pp. 134-138.

[26] R. Wang, C. Y. Chow and Y. Lyu, "Exploring cell tower data dumps for supervised learning-based point-of-interest prediction (industrial paper)", Geoinformatica, Springer, doi: 10.1007/s10707-015-0237-7, vol. 20, no. 2, pp. 327-349.