Facebook Comment Volume Prediction

Syed Faysal Kabir
Department of Computer Science
East West University
Dhaka, Bangladesh
kabirfaysal304@gmail.com

Md. Ridwan Sarker Turza
Department of Computer Science
East West University
Dhaka, Bangladesh
cse.turza@gmail.com

Md. Zakir Hossain Bhuiyan
Department of Computer Science
East West University
Dhaka, Bangladesh
zakirhossaunbhuiyan@gmail.com

***Abstract* — Data in the social networking services is increasing day by day. So, there is heavy requirement to study the highly dynamic behavior of the users towards these services. This work is a preliminary work to study and model the user activity patterns. We had targeted the most active social networking service 'Facebook' importantly the 'Facebook Pages' for analysis. The task here is to estimate the comment count that a post is expected to receive in next few hours. The analysis is done by modeling the comment patterns using variety of regressive modeling techniques. Additionally, we had also examined the effect of meta-learning algorithms over regression. For the whole analysis, a software prototype is developed consisting of (1) crawler, (2) pre-processor and (3) KDD module. After deep analysis, we conclude that the decision trees performed better than multi-layer preceptron neural networks. The effect of meta-learning algorithms is also inspected and it is visualized that the bagging had improved the results in terms of accuracy whereas dagging had improved the performance of the analysis.**

Keywords - Multi-Layer Preceptron (MLP); RBF Network; Prediction; Facebook; Comments; Data Mining; REP Tree; M5P Tree.

I. INTRODUCTION

The increasing use of social networking services had drawn the public attention explosively from last 15 years [1]. The merging up of physical things with the social networking services had enabled the conversion of routine objects into information appliances [2]. These services are acting like a multi-tool with daily applications like: advertisement, news, communication, banking, commenting, marketing etc. These services are revolutionizing day by day and many more on the way [3]. These all services have one thing in common that is daily huge content generation, that is more likely to be stored on hadoop clusters [4] [5]. As in Facebook, 500+ terabytes of new data ingested into the databases every day, 100+ petabytes of disk space in one of FB's largest Hadoop (HDFS) clusters and there is 2.5 billion content items shared per day (status updates + wall posts + photos + videos + comments). The Twitter went from 5,000 tweets per day in 2007 to 500,000,000 tweets per day in 2013. Flickr features 5.5 billion images as that of January 31,2011 and around 3k-5k images are adding up per minute [6].

In this research, we targeted the most active social networking service 'Facebook' importantly the 'Facebook Pages' for analysis. Our research is oriented towards the estimation of comment volume that a post is expected to receive in next few hours. Before continuing to the problem of comment volume prediction, some domain specific concepts are discussed below:

- *Public Group/Facebook Page:* It is a public profile specifically created for businesses, brands, celebrities etc.
- *Post/Feed:* These are basically the individual stories published on page by administrators of page.
- *Comment:* It is an important activity in social sites, that gives potential to become a discussion forum and it is only one measure of popularity/interest towards post is to which extent readers are inspired to leave comments on document/post.

To automate the process, we had developed a software prototype consisting of 3 major components, (1) crawler, (2) pre-processor and (3) KDD module. The crawler is a focused crawler and it crawls the Facebook pages of interest. The pre-processor module is responsible to pre-process the data and make it process ready and the KDD module is equipped with the number of *regression* modeling techniques for detailed analysis.

In the recent past, Singh K. et.al.[1], the authors had developed a software prototype demonstrating the comment volume prediction over Facebook pages using Neural Networks and Decision Trees and concluded that the Decision trees performed better than the Neural Networks. Buza K. [7], the authors had developed an industrial proofof-concept demonstrating the fine-grained feedback prediction on Hungarian blogs using various prediction models and on variety of feature sets and evaluated the results using Hits@10 and AUC@10 measures. Yano T. [8], the authors had modeled the relationship between content of political blog and the comment volume using Naive Bayes, Linear regression, Elastic regression and Topic-Poisson Models, and then evaluated them under the light of precision, recall and F1

measure. Rahman M.M. [9], had collected the different attributes such as about me, comments, wall post and age from facebook and analysed the mined knowledge with comparison to possible usages like as human behavior prediction, pattern recognition, job responsibility distribution, decision making and product promoting etc.

II. MACHINE LEARNING MODELS

The prediction process that is presented in this paper performs the comment volume prediction (CVP) using *regression* modeling technique. In this section, we discussed various examined regression techniques.

A. MLP

A multi-layer perceptron (MLP) is an artificial neural network model that maps the set of input data to the set of appropriate outputs. A MLP consist of multiple layers (Input layer, hidden layers, and output layers), of nodes that are connected in a directed fully connected graph, and with each layer fully connected to the next node. Each neuron (except the neurons in the first layer) in the artificial neural networks is equipped with a non-linear activation function. MLP make use of supervised learning technique for training the network [10] [11]. MLP is a modification to the standard linear perceptron and can distinguish data that are not separable directly [12].

B. REP Tree

Reduced error pruning tree (REP Tree) is a quick decision tree learner which builds a regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances [13] [14].

C. M5P Tree

M5P Tree [15] is a reconstruction of Quinlan's M5 algorithm [16] for inducing trees of regression models. M5P Tree combines the features of a conventional decision tree with linear regression functions at the nodes.

D. RBF Network

In the field of mathematical modeling, a radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters [17]-[19].

III. META-LEARNING ALGORITHMS

Meta learning is a process of learning to learn. These uses experience to change some aspects of learning algorithm or to the learning itself, so that the modified learning will be better than the original learning.

A. Bagging

Bootstrap aggregation or bagging [20] is a meta-learning technique designed to improve the accuracy and stability of machine learning algorithms.

It is a process of generating multiple statistical models by generating multiple training sets of same size by [1] bootstrap sampling. These multiple models are then aggregated to make a combined predictor by the process of voting and averaging. This way we can simulate the scenario of having multiple training sets. It can improve the performance of unstable learning model whereas the performance of a stable model can be deteriorated. It reduces the variance and helps to avoid the over fitting.

B. Dagging

Dagging or disjoint sample aggregation [21] is a metalearning technique. It creates multiple disjoint training set to train the regressors and then these multiple models are aggregated to make a combined predictor by the process of voting and averaging.

IV. PROBLEM FORMULATION

The task here is to estimate the comment count that a post is expected to receive in next few hours. Given some posts that appeared in past, whose target values (comments received) are already known, we simulated the scenario.

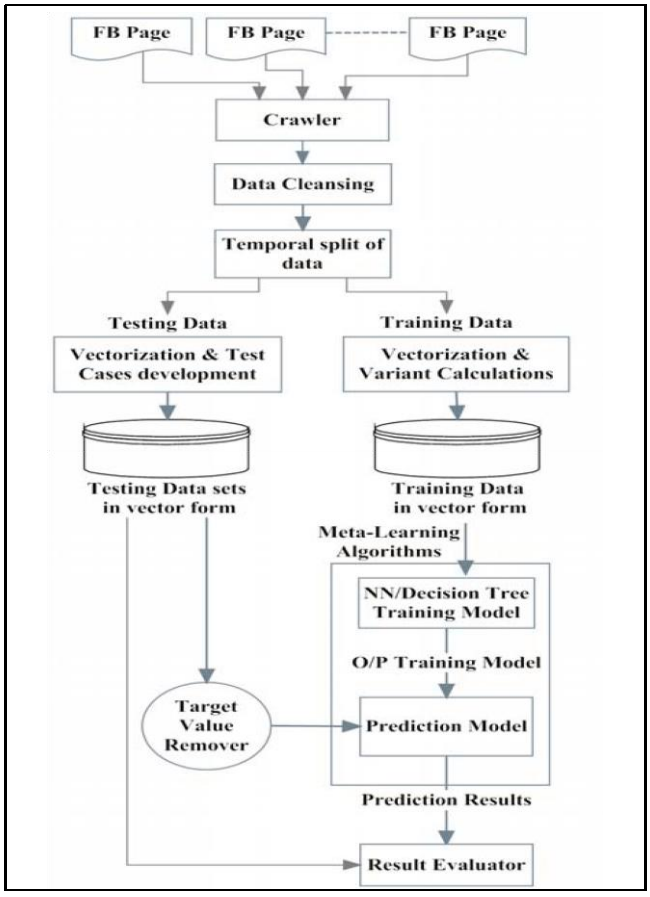

Figure 1. Process flow of the whole comment volume prediction process, starting from the Facebook pages crawling and ending to the prediction evaluation.

The analysis is done by modeling the comment patterns by using the variety of regressive modeling techniques. We

[1] It is a subset of training set that is developed randomly. The developed subset is smaller in size than original set.

address this problem as *regression* problem and various regression modeling techniques has been used to predict the comment volume.

Figure 1, demonstrates the process flow of the whole comment volume prediction process, starting from the Facebook pages crawling and ending to the prediction evaluation. For analysis, we crawled the Facebook pages for raw data, pre-processed it, and made a temporal split of the data to prepare the training and testing set. Then, this training set is used to train the regressor and performance of regressor is then estimated using testing data (whose target value is hidden) using some evaluation metrics.

A. Feature set used for this work

We had identified 53 features and 1 as target value for each post and categorized these features as:

1) Page Features: We identified 4 features of this category that includes features that define the popularity/Likes, category, checkin's and talking about of source of document. *Page likes:* It is a feature that defines users support for specific comments, pictures, wall posts, statuses, or pages. *Page Category*: This defined the category of source of document eg: Local business or place, brand or product, company or institution, artist, band, entertainment, community etc. *Page Checkin's:* It is an act of showing presence at particular place and under the category of place, institution pages only. *Page Talking About:* This is the actual count of users that were ‘ engaged’ and interacting with that Facebook Page. The users who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares by visitors to the page.

2) Essential Features: This includes the pattern of comment on the post in various time intervals w.r.t to the randomly selected base date/time demonstrated in Figure 2, named as C1 to C5.

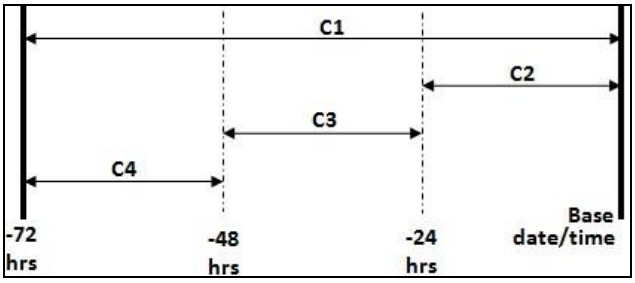

Figure 2. Demonstrating the essential feature details.

C1: Total comment count before selected base date/time. *C2:* Comment count in last 24 hrs with respect to selected base date/time. *C3:* Comment count is last 48 hrs to last 24 hrs with respect to base date/time. *C4:* Comment count in first 24 hrs after publishing the document, but before the selected base date/time. *C5:* The difference between C2 and C3. Furthermore, we aggregated these features by source and developed some derived features by calculating min, max, average, median and Standard deviation of 5 above mentioned features. So, adding up the 5 essential features and 25 derived essential features, we got 30 features of this category.

3) Weekday Features: Binary indicators *(0,1)* are used to represent the day on which the post was published and the day on selected base date/time. 14 features of this type are identified.

4) Other Basic Features: This include some document related features like length of document, time gap between selected base date/time and document published date/time ranges from *(0,71)*, document promotion status values *(0,1)* and post share count. 5 features of this category are identified.

B. Crawling

The data originates from Facebook pages. The raw data is crawled using crawler that is designed for this research work. This crawler is designed using JAVA and Facebook Query Language (FQL). The raw data is crawled by crawler and cleaned on basis of following criteria:

- We considered, only those comments that was published in last three days with respect to [2] base date/time as it is expected that the older posts usually don't receive any more attention.
- We omitted posts whose comments or any other necessary details are missing.

This way we produced the cleaned data for analysis.

C. Pre-processing

The crawled data cannot be used directly for analysis. So, it is carried out through many processes like split and vectorization. We made *temporal split* on this corpus to obtain training and testing data-set as we can use the past data(Training data) to train the model to make predictions for the future data(Testing data) [22] [23]. This is done by selecting a threshold time and divide the whole corpus in two parts. Then this data is subjected to *vectorization*. To use the data for computations it is required to transform that data into vector form. For this transformation, we had identified some features as already discussed in this section, on which comment volume depends and transformed the available data to vector form for computations. The process of vectorization is different in training and testing set:

1) Training set vectorization

In the training set, the vectorization process goes in parallel with the variant generation process. *Variant* is defined as, how many instances of final training set is derived from single instance/post of training set. This is done by selecting different base date/time for same post at random and process them individually as described in Figure 2. Variant - X, defines that, X instances are derived from single training instance as described in example of Facebook official page id: 103274306376166 with post id: 716514971718760, posted on Mon Aug 11 06:19:18 IST 2014, post crawled on Fri Aug 15 11:51:35 IST 2014. It received total of 515 comments at time of crawling as shown in Figure 3.

[2] Base date/time, It is selected to simulated the scenario, as we already know what will happen after this. There is one more kind of time we used in this formulation: is the post published time, which comes before the selected base date/time.

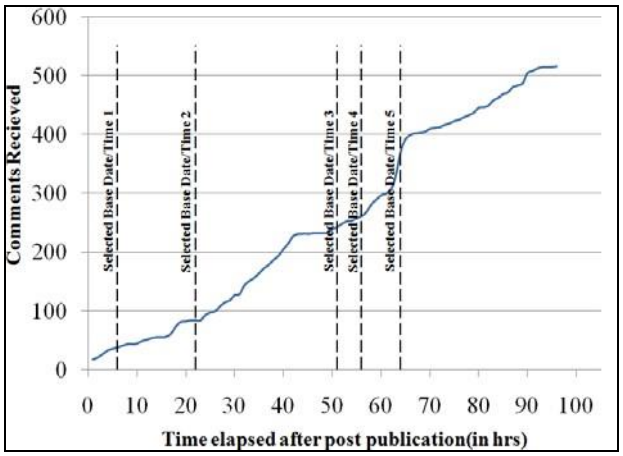

Figure 3. Cumulative Comments and different selected base date/time.

Know, by selecting different base date/time at random for single post, different variants are obtained for above example shown in Table 1.

TABLE I. VARIANTS OBTAINED

Variant	Selected Base Date/Time	Comments received in last 72 Hrs w.r.t Base Date/ Time	Comments target value
1	6	38	88
2	22	83	149
3	51	242	180
4	56	261	184
5	64	371	112

2) Testing set vectorization

Out of the testing set, 10 test cases are developed at random with 100 instances each for evaluation and then they are transformed to vectors.

D. Predictive Modeling

For the fine-grained evaluation, we have used the Decision Trees(REP Tree [13] [14] [24] and M5P Tree [15] [25] and Neural Networks(Multi-Layer Preceptron [10] [11] [26], RBF Network [17]-[19] [27] predictive modeling techniques.

1) Evaluation Metrics

The models and training set variants are evaluated under the light of Hits@10, AUC@10 and Evaluation Time as evaluation metrics:

2) Hits@10

For each test case, we considered top 10 posts that were predicted to have the largest number of comments, we counted that how many of these posts are among the top ten posts that had received the largest number of comments in actual. We call this evaluation measure *Hits@10* and we averaged Hits@10 for all cases of testing data [7]. Hits@10 is one of the important accuracy parameter for the proposed work. It tells about the prediction accuracy of the model.

3) AUC@10

For the AUC[28], i.e., area under the receiver-operator curve, we considered as positive the 10 blog pages receiving the highest number of feedbacks in the reality. It is represented as:

$$AUC = \frac{T_P}{T_P + F_P} \quad (1)$$

where, T_P is True positive's and F_P is False positives. AUC@10 metrics tells about the prediction precision of the models. Then, we ranked the pages according to their predicted number of feedbacks and calculated AUC. We call this evaluation measure AUC@10.

4) Evaluation time

It is the duration of the work performed describing the efficiency of the model. This measure includes the time to train the regressor and to evaluate the test cases.

V. EXPERIMENT SETTINGS

For our experiment, we crawled Facebook pages to collect the data for training and testing of our proposed model. In total 2,770 pages are crawled for 57,000 posts and 4,120,532 comments using JAVA and Facebook Query Language (FQL). The crawled data adds up to certain GB's and this process of crawling had taken certain weeks. After crawling, the crawled data is cleaned(After cleansing 5,892 posts are omitted and we left with 51,108 posts).

We divided the cleaned corpus into two subsets using temporal split, (1) Training data(80%, 40988) and (2) Testing data(20%, 10120) and then these datasets are sent to preprocessor modules for preprocessing where:

1) Training Dataset: The training dataset goes through a parallel process of variant calculations and vectorization and as a result of training set pre-processing, we are obtained with these five training sets as:

TABLE II. TRAINING SET VARIANTS.

Training set Variant	Instance Count
Variant - 1	40,949
Variant - 2	81,312
Variant - 3	121,098
Variant - 4	160,424
Variant - 5	199,030

2) Testing Dataset: Out of 10,120 testing data items, 1000 test posts are selected at random and 10 test cases are developed are described earlier.

The models that are used for experiments are Multi-Layer preceptron(MLP), RBF Networks, Decision Trees(REP Tree and M5P Tree). We used *WEKA* (The Waikato Environment for Knowledge Analysis) implementations of these regressors.

Neural Network - Multi Layer Perceptron Learning is used in 2 forms: (1)Single Hidden layer with 4 neurons. and (2) two hidden Layers, 20 neurons in 1 hidden layer and 4 in 2nd hidden layer. For both of the cases, the training iterations are fixed to 100, while the learning rate to 0.1 and momentum to 0.01. For Radial Basial function (RBF) Network, the cluster count is set to 90 clusters and default parameters are used for REP and M5P Tree.

VI. RESULT AND DISCUSSION

The experimentation had been performed on variety of regression models and variety of datasets. Table 2, presents the results of Hits@10, AUC@10 and Evaluation time, without any meta-learning algorithm.

TABLE III. EXPERIMENTAL RESULTS

Model		Variant – 1	Variant – 2	Variant – 3	Variant – 4	Variant – 5
MLP – 4	Hits@10	5.500 ± 1.285	6.200 ± 1.166	6.200 ± 0.980	5.800 ± 1.661	5.700 ± 1.345
	AUC@10	0.656 ± 0.164	0.807 ± 0.189	0.852 ± 0.180	0.795 ± 0.232	0.670 ± 0.205
	Time Taken	40.882 Sec	190.809 Sec	132.469 Sec	162.377 Sec	193.465 Sec
MLP – 20,4	Hits@10	5.300 ± 1.345	6.300 ± 1.187	6.200 ± 0.980	6.400 ± 1.114	5.700 ± 1.005
	AUC@10	0.674 ± 0.157	0.831 ± 0.193	0.809 ± 0.206	0.832 ± 0.190	0.734 ± 0.205
	Time Taken	166.804 Sec	335.025 Sec	474.729 Sec	629.820 Sec	777.803 Sec
REP Tree	Hits@10	5.900 ± 1.640	6.000 ±1.000	6.400 ± 0.917	6.700 ± 1.187	6.600 ± 1.281
	AUC@10	0.784 ± 0.127	0.827 ± 0.121	0.768 ± 0.109	0.807 ± 0.098	0.756 ± 0.137
	Time Taken	10.844 Sec	9.885 Sec	28.618 Sec	41.483 Sec	46.871 Sec
M5P Tree	Hits@10	6.100 ± 1.300	6.700 ± 0.900	6.000 ± 1.183	6.300 ± 0.781	6.100 ± 1.578
	AUC@10	0.761 ± 0.143	0.708 ± 0.165	0.711 ± 0.165	0.693 ± 0.199	0.730 ± 0.185
	Time Taken	34.440 Sec	71.520 Sec	117.599 Sec	177.850 Sec	518.638 Sec
RBF Network (90 Clusters)	Hits@10	4.100 ± 1.136	4.500 ± 1.025	4.100 ± 1.221	3.300 ± 1.345	3.600 ± 1.428
	AUC@10	0.899 ± 0.110	0.912 ± 0.087	0.945 ± 0.083	0.937 ± 0.077	0.912 ± 0.086
	Time Taken	298.384 Sec	491.002 Sec	614.138 Sec	1602.836 Sec	1831.946 Sec

1) Hits@10

From the graph shown in Figure 4, it is observed that the prediction Hits@10 accuracy in case of decision trees is higher compared to other modeling techniques and RBF Model shown minimal Hits@10 accuracy.

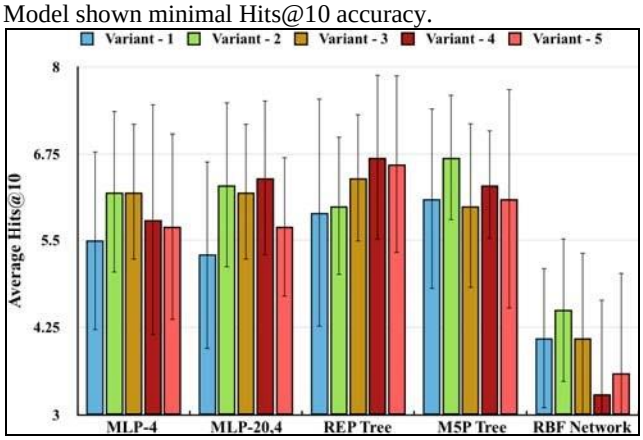

Figure 4. Hits@10 for comment volume prediction is presented in this graph along with the standard deviation.

The Hits@10 measure of REP Tree is *6.700 ± 1.187* of dataset variant - 4 and M5P Tree is *6.700 ± 0.900* of dataset variant - 2. Whereas, it is minimal in case of RBF Network, that is *3.300 ± 1.345* of dataset variant - 4.

2) AUC@10

From the graph shown in Figure 5,

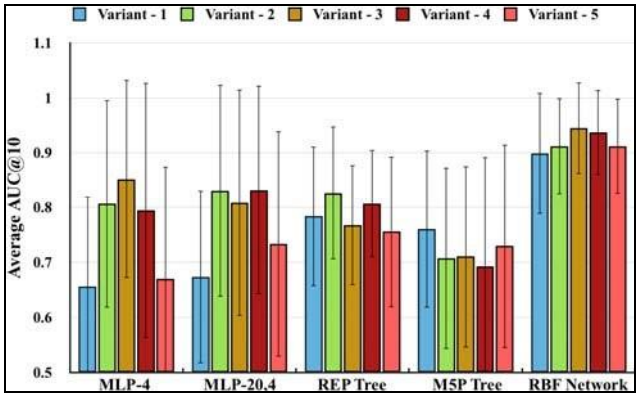

Figure 5. AUC@10 for comment volume prediction is presented in this graph along with the standard deviation.

It is observed that the prediction precision i.e: AUC@10 in case of RBF Network is higher compared to the other prediction models used. The RBF Network have maximum prediction precision of *0.945 ± 0.083* of dataset variant - 3 and minimum in the case of MLP of 1 hidden layer with 4 neurons that is *0.656 ± 0.164* of variant - 1.

3) Evaluation Time

From the graph in Figure 6, it is observed that the Evaluation time is minimal in the case of REP Tree and maximum in the case of RBF network. It is also observed that the evaluation time is directly proportional to the variant size.

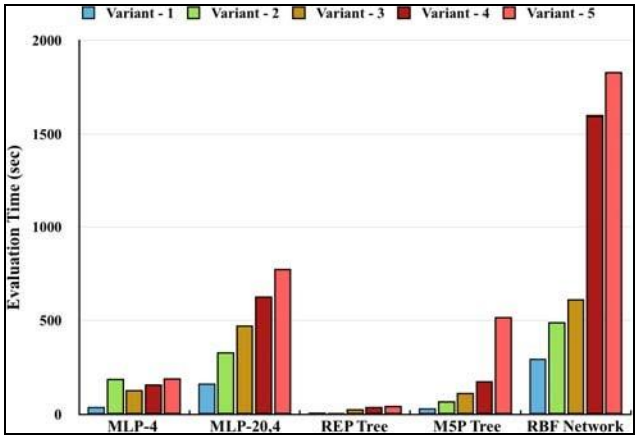

Figure 6. Evaluation time for comment volume prediction is presented in this graph.

Through the deep analysis of the prediction results, it is observed that the decision trees have better accuracy and precision under the light of all evaluation metrics. It is also observed that the evaluation time is maximum in case of RBF Network.

A. Bagging

The effect of bagging has been measured on the comment volume prediction process. Table 4, presents the results of comment volume prediction when bagging meta learning algorithm is used for analysis.

1) Hits@10

From the graph shown in Figure 7 and Figure 4, it is observed that the accuracy of the prediction has been increased for all analyzed models using bagging.

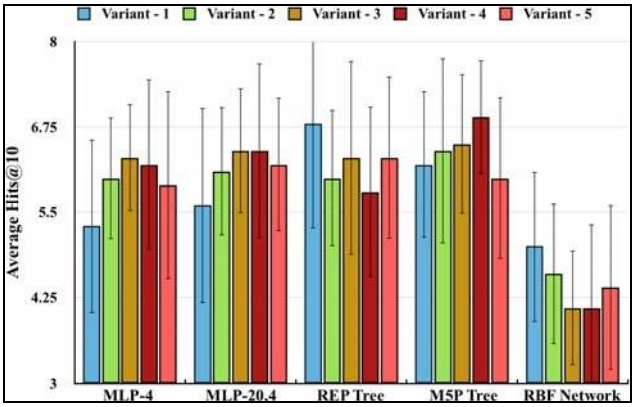

Figure 7. Hits@10 for comment volume prediction is presented in this graph along with the standard deviation.

For REP tree the value is increased from *5.900 ± 1.640* of variant - 1 to *6.800 ± 1.536* and of M5P tree the value is increased from *6.300 ± 0.931* of variant - 4 to *6.900 ± 0.831*. For the neural network, bagging had improved the accuracy like in MLP-4 the value is increased from *5.800 ± 1.661* of variant - 4 to *6.200 ± 1.249*. The RBF network had shown similar effect is on prediction.

TABLE IV. EXPERIMENTAL RESULTS – BAGGING

Model		Variant – 1	Variant – 2	Variant – 3	Variant – 4	Variant – 5
MLP – 4	Hits@10	5.300 ± 1.269	6.000 ± 0.894	6.300 ± 0.781	6.200 ± 1.249	5.900 ± 1.375
	AUC@10	0.681 ± 0.134	0.781 ± 0.227	0.833 ± 0.183	0.836 ± 0.204	0.761 ± 0.194
	Time Taken	360.626 Sec	723.550 Sec	1067.561 Sec	1574.212 Sec	1782.111 Sec
MLP – 20,4	Hits@10	5.600 ± 1.428	6.100 ± 0.943	6.400 ± 0.917	6.400 ± 1.281	6.200 ± 0.980
	AUC@10	0.735 ± 0.147	0.816 ± 0.198	0.811 ± 0.197	0.828 ± 0.186	0.792 ± 0.177
	Time Taken	1576.265 Sec	3138.339 Sec	5065.178 Sec	6462.098 Sec	7773.225 Sec
REP Tree	Hits@10	6.800 ± 1.536	6.000 ± 1.000	6.300 ± 1.418	5.800 ± 1.249	6.300 ± 1.187
	AUC@10	0.746 ± 0.122	0.727 ± 0.113	0.761 ± 0.121	0.634 ± 0.128	0.722 ± 0.098
	Time Taken	52.722 Sec	137.485 Sec	229.443 Sec	302.766 Sec	412.973 Sec
M5P Tree	Hits@10	6.200 ± 1.077	6.400 ± 1.356	6.500 ± 1.025	6.900 ± 0.831	6.000 ± 1.183
	AUC@10	0.781 ± 0.148	0.714 ± 0.164	0.754 ± 0.181	0.610 ± 0.165	0.812 ± 0.156
	Time Taken	318.466 Sec	655.642 Sec	1142.914 Sec	1411.762 Sec	1922.949 Sec
RBF Network (90 Clusters)	Hits@10	5.000 ± 1.095	4.600 ± 1.020	4.100 ± 0.831	4.100 ± 1.221	4.400 ± 1.200
	AUC@10	0.964 ± 0.047	0.939 ± 0.076	0.949 ± 0.064	0.945 ± 0.064	0.945 ± 0.063
	Time Taken	2445.239 Sec	4619.697 Sec	10171.881 Sec	11996.384 Sec	16344.734 Sec

2) AUC@10

From the graph shown in Figure 8 and Figure 5, It is observed that the precision is increased for all models using bagging like in case of RBF network the precision is increased from *0.899 ± 0.110* to *0.964 ± 0.047* for variant –

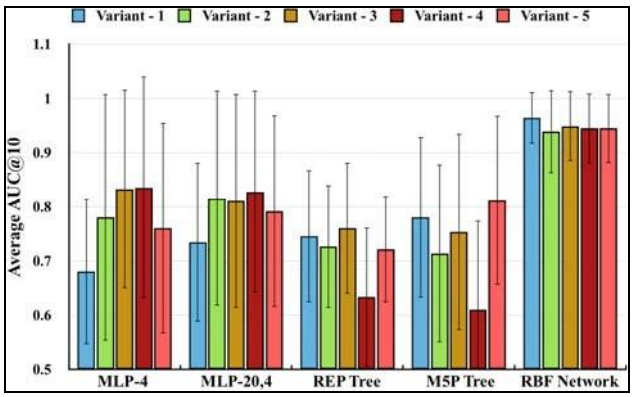

Figure 8. AUC@10 for comment volume prediction is presented in this graph along with the standard deviation.

3) Evaluation Time

From the graph in Figure 9 and Figure 6, it is observed that the prediction time is increased for all models using The evaluation time is directly proportional to the training bagging. dataset size like minimal time is required for variant - 1 and

There is a uniform increase in the prediction time using bagging like in case of REP Tree the prediction time is increased from *10.844 Sec* to *52.722 Sec* for variant - 1 and from *46.871 Sec* to *412.973 Sec* for variant - 5.

1. For *MLP-(20,4)*, the precision is increased from *0.734 ± 0.205* to *0.792 ± 0.177* for variant - 5. Whereas, in case of decision trees the bagging had shown very little variation on prediction, like AUC is increased from *0.730 ± 0.185* to *0.812 ± 0.156* for variant - 5, and is decreased from *0.807 ± 0.098* to *0.634 ± 0.128* for variant - 4.

maximum time is for

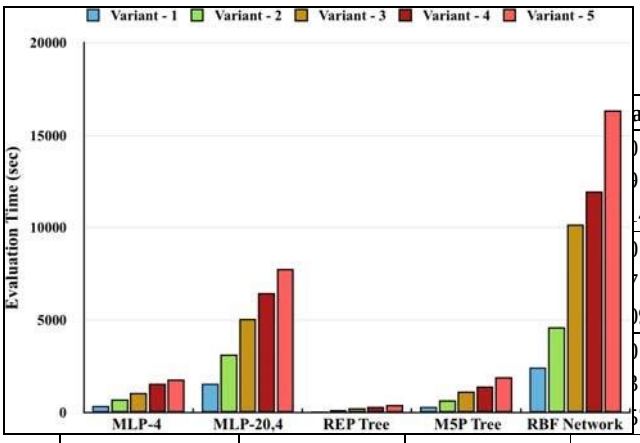

Figure 9. Evaluation time for comment volume prediction is presented in this graph.

		ant – 2	V
		) ± 1.676	4
		) ± 0.176	0
		.4 Sec	1
		) ± 1.628	4
		' ± 0.151	0
		098 Sec	4
		) ± 1.269	6
		3 ± 0.092	0
		57 Sec	1
M5P Tree	Hits@10	6.000 ± 0.775	6.400 ± 0.917	6
	AUC@10	0.814 ± 0.140	0.808 ± 0.141	0
	Time Taken	21.616 Sec	41.800 Sec	7
RBF Network (90 Clusters)	Hits@10	4.600 ± 0.800	4.100 ±1.375	4
	AUC@10	0.846 ± 0.290	0.714 ± 0.323	0
	Time Taken	145.172 Sec	322.602 Sec	6

B. Dagging

The effect of dagging has been measured on the comment volume prediction process. Table 5, presents the results of comment volume prediction when dagging meta learning algorithm is used for analysis.

1) Hits@10

From the graph in Figure 10 and Figure 4, It is observed that the prediction accuracy of neural networks and decision trees is deteriorated by using dagging meta learning algorithm and of RBF network, accuracy is improved.

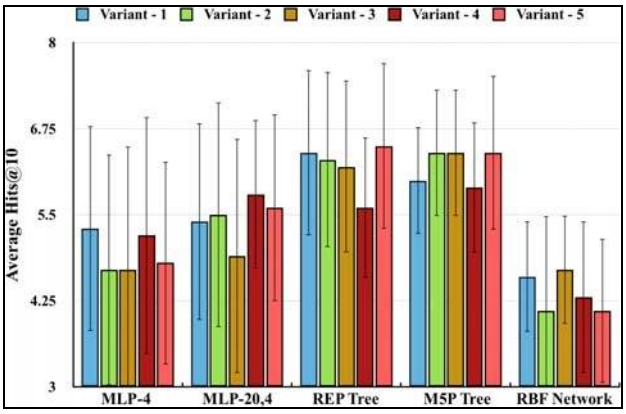

Figure 10. Hits@10 for comment volume prediction is presented in this graph along with the standard deviation.

Like in case of REP tree the Hits@10 value is decreased from *6.700 ± 1.187* to *5.600 ± 1.020* for variant - 5. In case of M5P tree the value is decreased from *6.700 ± 0.900* to *6.400 ± 0.917* for variant - 2. Whereas, in case of RBF network the value is increased for most datasets.

2) AUC@10

From the graph shown in Figure 11 and Figure 5,

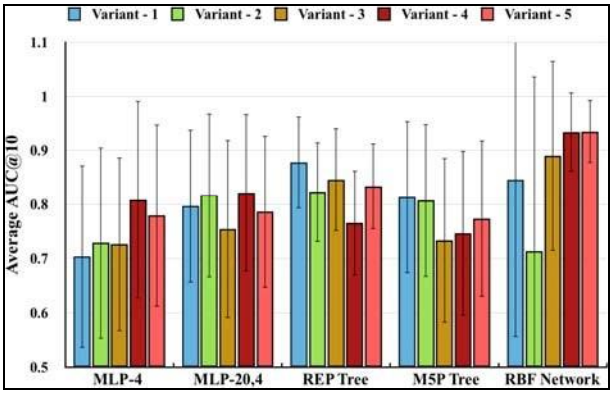

Figure 11. AUC@10 for comment volume prediction is presented in this graph along with the standard deviation.

It is observed that by using dagging meta learning algorithm the prediction precision of decision trees is increased, whereas for neural networks it is deteriorated.

3) Evaluation Time

From the graph in Figure 12 and Figure 6, it is observed that the prediction time is decreased for all models using dagging.

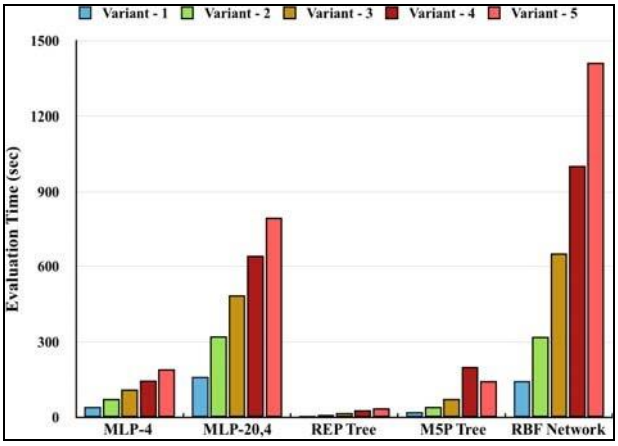

Figure 12. Evaluation time for comment volume prediction is presented in this graph.

In case of REP tree the evaluation time is decreased from *10.844 Sec* to *6.034 Sec* for variant - 1 and from *46.871 Sec* to *37.627 Sec* for variant - 5. For M5P tree it again decreased from *34.440 Sec* to *21.616 Sec* for variant - 1 and from *518.638 Sec* to *146.027 Sec* for variant - 5. This shows that the evaluation time is directly proportional to the training dataset.

VII. Conclusion and Future Scope

We examined the neural network and decision trees using a set of training data sets and came to the conclusion that the highly dynamic user behaviour can be modelled to make the future estimations. In our analysis, we used the decision trees and neural networks and found that the decision trees perform better than the neural networks for this comment volume prediction process. Another study that is made in this paper is to measure the effect of metalearning algorithms and we found that the bagging meta learning algorithm had improved the prediction accuracy whereas the evaluation time is higher in this case and dagging meta-learning algorithm had improved the prediction performance whereas it had deteriorated the performance of the prediction.

The outcome of this work is a software prototype for comment volume prediction which can be further enhanced by using (1) category based predictor , (2) by including multi-media features, (3) by using a hybrid set of regressors for better modeling or by using metaheuristic modelling techniques.

Acknowledgment

The authors would like to thank Facebook for providing the necessary API's for data crawling, without which the proposed work was not feasible. This manuscript is an extended manuscript of the paper entitled “Comment Volume Prediction using Neural Networks and Decision Trees”, presented at IEEE 2015 17th UKSIM-AMSS International Conference on Modeling and Simulation, UKSim2015, Cambridge University, Cambridge, United Kingdom *DOI 10.1109/UKSim.2015.20*.

References

[1] 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, Mar. 2015.

[2] A. Kamilaris and A. Pitsillides, "Social networking of the smart home," in Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on, Sept 2010, pp. 2632–2637.

[3] Y. Meguebli, M. Kacimi, B.-l. Doan, and F. Popineau, "How hidden aspects can improve recommendation?" in Social Informatics, ser. Lecture Notes in Computer Science, L. Aiello and D. McFarland, Eds. Springer International Publishing, 2014, vol. 8851, pp. 269–278.

[4] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, May 2010, pp. 1–10.

[5] I.Polato,R.Re´ ,A.Goldman,andF.Kon,"Acomprehensiveviewof hadoop researcha systematic literature review," Journal of Network and Computer Applications, vol. 46, pp. 1–25, 2014.

[6] T. Reuter, P. Cimiano, L. Drumond, K. Buza, and L. Schmidt-Thieme, "Scalable event-based clustering of social media via record linkage techniques." in ICWSM, 2011.

[7] K. Buza, "Feedback prediction for blogs," in Data Analysis, Machine Learning and Knowledge Discovery, ser. Studies in Classification, Data Analysis, and Knowledge Organization, M. Spiliopoulou, L. Schmidt-Thieme, and R. Janning, Eds. Springer International Publishing, 2014, pp. 145–152.

[8] T. Yano and N. A. Smith, "What's worthy of comment? content and comment volume in political blogs." in ICWSM, 2010.

[9] M. M. Rahman, "Intellectual knowledge extraction from online social data," in Informatics, Electronics & Vision (ICIEV), 2012 International Conference on. IEEE, 2012, pp. 205–210.

[10] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," DTIC Document, Tech. Rep., 1961.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep., 1985.

[12] G. Cybenko, "Approximation by superpositions of a sigmoidal function," Mathematics of control, signals and systems, vol. 2, no. 4, pp. 303–314, 1989.

[13] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," Advances in Space Research, vol. 41, no. 12, pp. 1955–1959, 2008.

[14] W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A comparative study of reduced error pruning method in decision tree algorithms," in Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on. IEEE, 2012, pp. 392–397.

[15] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," 1996.

[16] J. R. Quinlan et al., "Learning with continuous classes," in 5th Australianjointconferenceonartificialintelligence,vol.92. Singapore, 1992, pp. 343–348.

[17] D. S. Broomhead and D. Lowe, "Radial basis functions, multivariable functional interpolation and adaptive networks," DTIC Document, Tech. Rep., 1988.

[18] D. Lowe, "Multi-variable functional interpolation and adaptive networks," Complex Systems, vol. 2, pp. 321–355.

[19] F. Schwenker, H. A. Kestler, and G. Palm, "Three learning phases for radial-basis-function networks," Neural networks, vol. 14, no. 4, pp. 439–458, 2001.

[20] L.Breiman,"Baggingpredictors,"MachineLearning,vol.24,no.2, pp. 123–140, 1996.

[21] K. M. Ting and I. H. Witten, "Stacking bagged and dagged models," in Fourteenth international Conference on Machine Learning, D. H. Fisher, Ed. San Francisco, CA: Morgan Kaufmann Publishers, 1997, pp. 367–375.

[22] T. M. Pelusi, D., "Optimal trading rules at hourly frequency in the foreign exchange markets," Mathematical and Statistical Methods for Actuarial Sciences and Finance, Springer, pp. 341–348, 2012, iSBN: 978-88-470-2341-3.

[23] D. Pelusi, M. Tivegna, and P. Ippoliti, "Improving the profitability of technical analysis through intelligent algorithms," Journal of Interdisciplinary Mathematics, vol. 16, no. 2-3, pp. 203–215, 2013.

[24] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," Advances in Space Research, vol. 41, no. 12, pp. 1955–1959, 2008.

[25] E. Onyari and F. Ilunga, "Application of mlp neural network and m5p model tree in predicting streamflow: A case study of luvuvhu catchment, south africa," in International Conference on Information and Multimedia Technology (ICMT 2010), Hong Kong, China, 2010, pp. V3–156.

[26] D. Pelusi, "Designing neural networks to improve timing performances of intelligent controllers," Journal of Discrete Mathematical Sciences and Cryptography, vol. 16, no. 2-3, pp. 187–193, 2013.

[27] A. G. Bors, "Introduction of the radial basis function (RBF) networks," in Online symposium for electronics engineers, vol. 1, no. 1, 2001, pp. 1–7.

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