ISSN (P): 2412-0448, ISSN (E): 1998-4154

A Comprehensive Comparative Evaluation of Machine Learning Algorithms on Facebook Comment Dataset

Iffraah Rehman
Institute of Business Management, Pakistan
iffraahrehman@gmail.com

Muhammad Umair
Universiti Teknologi PETRONAS, Malaysia,
muhammad_17008606@utp.edu.my

Shamim Akhtar
Majmaah University, Saudi Arabia
shamimakhtar1@gmail.com

Waqar Khan
Institute of Business Management, Pakistan,
waqar.khan@iobm.edu.pk

Haider Abbas
Iqra University, Pakistan
haider@iqra.edu.pk

Ranjay Kumar Choudhary
Majmaah University, Saudi Arabia,
r.choudhary@mu.edu.sa

Abstract**-- Data mining is an emerging technique with its application in various areas such as health care, education, travel, social media, and banking. The data can be either labeled or unlabeled. When it comes to social media, the various platforms generate an infinite amount of data. This data can be of immense importance as a lot of hidden information can be discovered after data mining. In this paper, machine-learning algorithms such as Decision Trees, SVM, and Linear Regression and their variants are applied on the Facebook comment dataset, obtained from the UCI machine learning repository. The dataset has 40,949 instances and 54 attributes. The goal is to predict the number of comments a Facebook post will get based on various conditions. The results indicate that the Fine Gaussian SVM variation of SVM yielded the highest prediction accuracy. The evaluation was done on different parameters such as average testing accuracy (%), Root Mean Square Error (RMSE), R- Squared, Mean Square Error (MSE), Mean Absolute Error (MAE), prediction speed (Obs/sec), and training time (Machine cycle). It is concluded that SVM is an ideal choice to solve prediction problems associated with social media data.**

Index Terms-- Data mining; Linear Regression; Root Mean Square Error; Social media data

I. INTRODUCTION

Social media platforms are now taking advantage of the latest developments in the field of machine learning to improve and elevate their services. Facebook, the social media giant, has been using and developing algorithms for text and image analysis, recommendation engines, and much more [1]. Some of the data from social media is publicly available, which has allowed researchers to analyze it using machine learning algorithms [2], for the purpose of classification and regression. These algorithms are applied by performing training and testing on datasets, and then being used for predictions, pattern recognition, and semantic analysis. Machine learning is divided into three categories, i.e., supervised, unsupervised, and reinforcement learning [3]. Since social networking sites have been helping towards the improvement in daily life, they can be used in the fields of medicine, business, and education [4]. Previously, it was difficult to collect the data from social media sites. It was done by conducting surveys of small sample space [5]. As of now, more content is available on social media, which is easily accessible. The data extracted from social media is then used for text classification, link mining, subgroup detection, etc. Social media are the online platforms used for communication, interaction, and information sharing. Millions of people from all over the world are connected through social media, resulting in a massive amount of data being posted and available online. Platforms such as Facebook and Twitter are the most used social networks for communication and information sharing [6]. Any tweet or post made by a user on such platforms is instantly shared with other users. Positive and negative feedback on the post can be seen, which sometimes analyzed and important decisions made.

II. RELATED WORK

A lot of work has been done over the last few years on the available set of social media, as well as the application of machine learning algorithms on that data. In paper [3], the author has predicted the financial performance of a company that deals with cars. The data was extracted from Twitter using hashtags for cars such as BMW and Volkswagen. The tool that has been used in WEKA, and algorithms applied are Random Forest, Ada Boost and Naïve Bayes. The main goal was to predict the Return on Assets (ROA). It was discovered that 86.17% of tweets can predict whether the company will perform well or poorly in the upcoming fiscal quarter. The author of the paper [7] has predicted the number of comments a Facebook post would receive, by using the comparative analysis of three algorithms such as Multilayer Perceptron (MLP), RBF Networks, and Decision Tress (REP and M5P Trees). WEKA has been used for the application of algorithms on the available dataset. The parameters to identify the best algorithm among all were Hits@10, auc@10, time taken and mean absolute error (MAE). However, the decision tree was found to be the best for this type of prediction purpose. Data mining plays and important role in the identification of important factors that are affecting Facebook usage in a positive and negative way. The authors in paper [8] analyzed the data collected from 570 active users of Facebook from Anakara. Algorithms such as Decision Tree, SVM, and Artificial Neural Network were applied to the dataset. The goal was to identify the major factors affecting the access frequency and usage time of Facebook. In this regard, SVM was found to give the most accurate results of all. This survey [14] discusses the use of social media. Social media sites are now part of all users' daily lives, mostly of the younger generation, as they are used in many ways for socializing, education, and communication, etc. In this study, some usage patterns, which were the usage of time and access frequency of Facebook users, were mined. The Decision Tree method was used for mining the data because of its ease of use and visual representation capabilities. It was concluded that SVM gave the most accurate results. It was also found that Facebook plays a vital role in communication between students and teachers, which makes Facebook a huge platform that contains different types of information and resources. In a survey by Asur and Huberman [15] it is discussed that how the use of social media is helping us in predicting the future. It was discussed that the users have become increasingly reliant on social media for networking and content sharing. In this case, 3 million tweets were analyzed and a chatter rate was calculated. In this paper, a linear regression model is used to forecast the revenue of yet-to-be-released films. To improve the prediction, sentiments were also analyzed in relation to the tweets in this paper.

III. MACHINE LEARNING ALGORITHMS

A. *Decision Trees*

One of the most basic supervised machine learning algorithms is the decision tree. It is simple to comprehend for humans. It is based on the divide-and-conquer strategy. The tree begins with a root which has branches based on rules. Then leaf has the decision outcome. The training sets are divided into subsets [9]. It can also deal with missing and redundant values, and it is resistant to noisy data. It also costs less to generate a model than other algorithms [10].

B. *Support Vector Machine*

The Support Vector Machine (SVM) is another machine learning technique that is used for unsupervised data. This algorithm employs a hyperplane to prevent data misclassifications by establishing data boundaries. It has no trouble dealing with underlying or hidden data, however, it has a major drawback in that it takes a long time to train, and it requires data with fewer instances of data. On the other hand, it can efficiently handle categorical and multivariate data [11].

C. *Linear Regression Model*

Another well-known algorithm in data mining is the linear regression algorithm. As the name implies, this model is based on the linear equation relationship between two variables. In a linear relationship, X is treated as a single input, while Y is treated as a single output, and Y is dependent on the X variable. The criterion is the dependent variable, and the predictor is the other variable. Linear regression is represented in a very straightforward manner. The best fit straight line is the graph's straight line.

D. *Ensemble of Trees*

Ensemble learning necessitates that the base models be as accurate and error-free as possible. No technique other than decision, on the other hand, can provide ensemble learning because it provides an accurate way of generating hypotheses. Additionally, bagging and boosting sampling techniques are used in conjunction with decision trees [12]

IV. EXPERIMENTAL ANALYSIS

A. *Description of Dataset*

The dataset used for this paper is obtained from UCI machine learning repository. The name of dataset in Facebook Comment Volume Dataset Data Set [URL:https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset]. The features in this dataset were extracted from various Facebook posts from the year 2016.

It has 54 attributes in total, out of which 53 are considered to be predictors, while the last attribute namely ‘No. of Comments’ is considered to be response. Total instances in the dataset are 40949, with less or no missing values. However, the data is divided into training and testing for the application of different algorithms. Table 1 lists the name of all the attributes that were used in the dataset along with their description.

Table 1 Attributes of the Dataset

Attribute No.	Attribute Name	Description
1	Page Popularity/Likes	Defines the support for the post of Facebook
2	Page Checking	Describes how many people has visited the place e.g. place, institution etc.
3	Page Talking about	Defines the activity on a particular post such s comments, likes, shares etc.
4	Page Category	Category of source post place, institution, brand
5-29	Derived	These features are aggregated by page, gives the min, max, average, median and standard deviation of standard features
30	CC1	Total number of comments before the selection of base time
31	CC2	Total number of comments in last 24 hours, relative to base time
32	CC3	Total number of comments in last 24 to 48 hours, relative to base time
33	CC4	The number of comments in the first 24 hours after the publication of post but before base date/time.
34	CC5	The difference between CC2 and CC3
35	Base time	Selected time to simulate the order
36	Post length	Number of characters in the post
37	Post Share Count	Number of times the post was shared
38	Post promotion Status	Represents the post was promoted (1) or not (0)
39	H Local	Describes the H hours for which the comments will be received
40-46	Post Published Weekday	Represents the weekday on which the post was published
47-53	Base Date Time Weekday	Represents the weekday on selected base time and date
54	Target Variable	The number of comments on post in next H hours

B. Methods and Materials

The purpose of this paper is to compare regression algorithms. The above-mentioned dataset is subjected to these algorithms. The tool that was used for this was MATLAB version 2018. The regression learner app is provided by MATLAB. This app contains all of the regression algorithms as well as various models for each algorithm. Several parameters are used to compare algorithms in order to determine which algorithm is the best fit for the dataset. The parameters used for comparison are as follows:

- Average testing accuracy (%)
- Root Mean Square Error (RMSE)
- R-Squared
- Mean Absolute Error (MAE)
- Mean Square Error (MSE)
- Prediction Speed (obs/sec)
- Training Time (sec)

The testing accuracy is the most important parameter of a regression algorithm. Fine Gaussian SVM algorithm shows 100 percent accuracy in the selected dataset. On the other hand, the Interactions linear algorithm has the lowest accuracy of all, at 69.94 percent. Furthermore, when the parameters of all algorithms' errors were analyzed, the Fine Gaussian SVM was found to have the lowest error rate, with R-squared equal to 0.04, RMSE equal to 36.292, and MAE equal to 6.074. It also has a prediction speed of 600 observations per second.

C. Table of Results

Table 2 shows a tabular comparison of all of the algorithms used on the dataset, as well as their parameters. The testing accuracy of the Fine Gaussian SVM algorithm is 100 percent with fewer errors. Another linear regression algorithm, robust linear, had an accuracy of approximately 93.21 percent, making it the second most accurate testing algorithm on the list. It was also discovered that none of the decision tree algorithms could produce satisfactory results. The average testing accuracy for all three algorithms, Fine, Medium, and Coarse, is nearly 75 percent. It can be said that algorithms classified as SVM are reliable for the dataset in question.

Table 2 Comparative Table of Regression Algorithms

	Root Mean Square Error (RMSE)	**R-Squared**	**Mean Squared Error (MSE)**	**Mean Absolute Error (MAE)**	**Prediction Speed (K obs/sec)**	**Training Time (sec)**	**Average Testing Accuracy %**
Linear Regression	31.957	0.25	1021.3	5.408	880	183.3	78.2457
Interactions Linear Regression	24.046	0.58	578.21	6.5199	27000	88.369	69.9423
Robust Linear	32.884	0.21	1081.4	5.513	610000	14.289	93.2198
Fine Tree	14.11	0.85	199.11	2.5261	57000	4.3761	75.2091
Medium Tree	19	0.74	361	3.4498	1200000	5.4247	75.2091
Coarse Tree	23.122	0.61	534.63	4.1457	1100000	3.8424	76.3441
Linear SVM	31.964	0.25	1021.7	5.4089	890	175.58	90.6113
Quadratic SVM	27.171	0.46	738.27	4.4138	790	1427.3	84.0402
Fine Gaussian SVM	36.292	0.04	1317.1	6.074	600	183.97	100
Medium Gaussian SVM	33.561	0.18	1126.4	5.394	800	160.47	85.8921
Coarse Gaussian SVM	33.011	0.2	1089.7	5.4992	720	167.87	92.6025
Boosted Tree	18.094	0.76	327.4	4.0421	84000	36.955	71.147
Bagged Tree	17.315	0.78	299.82	2.9885	35000	22.473	70.4401

D. Testing Graphs

The testing graphs of all the applied algorithms are shown in the following figures. It is possible to see the difference between actual values and values predicted by the algorithm after training. The red coolers in the graph represent the predicted values, while the blue colors represent the actual values in the response column. The testing graph of the Fine Tree algorithm is shown in Figure 1. The accuracy of the testing was found to be 74.89 percent. The algorithm was unable to predict the actual number of comments that a given post would receive. Figure 2 depicts the accuracy of testing achieved by using Medium Tree. Based on the various attributes, Medium Tree was only able to predict 75.20 percent of the correct numbers of comments a Facebook would receive. The graphical representation of testing accuracy of Coarse Tree is shown in Figure 3. The algorithm was able to predict a 76.34 percent average percentage of correct output. Predicting the response column values took the longest time, at 1100000 obs/sec. Despite having a shorter prediction time and a lower number of errors, linear regression failed to predict the correct output as required. The testing accuracy of linear regression is illustrated as 78.24 percent in Figure 4. The interaction linear algorithm is used to test the response column in Figure 5, and the algorithm produces the worst results. It has the lowest testing accuracy of any of the tests, at only 69.94%. The testing accuracy of the robust linear algorithm was found to be 93.21 percent. As shown in Figure 6, only a few of the values in the response column were predicted by the algorithm to be the same as the actual ones. As shown in Figure 7, Linear SVM was able to correctly predict approximately 90.61 percent of the values. The remaining 10%, on the other hand, were found to have predicted values that were significantly higher than the actual ones. Figure 8 shows that Quadratic SVM predicted some of the values to be negative, even though the response column should not have any negative values. This algorithm's actual testing accuracy is calculated to be 84.04 percent. As shown in Figure 9, the Fine Gaussian SVM has a testing accuracy of 100%. Indicates that the predicted and actual values were identical. While the model was being trained for the algorithm, it had the fewest errors. Among all the algorithms used in this study, this is the best. Figure 10 depicts the testing accuracy obtained by using a medium Gaussian SVM. Based on the various attributes, Medium Tree was only able to predict 85.8921 percent of the correct number of comments a Facebook would receive. Like other variation of decision trees, the ensemble of trees such as bagged tree in the Figure 11shows the average testing accuracy of 70.44%. Figure 12 shows the accuracy of the testing obtained by using Boosted Tree. Based on the various attributes, Boosted Tree could only predict 70.44 percent of the correct number of comments a Facebook would receive.

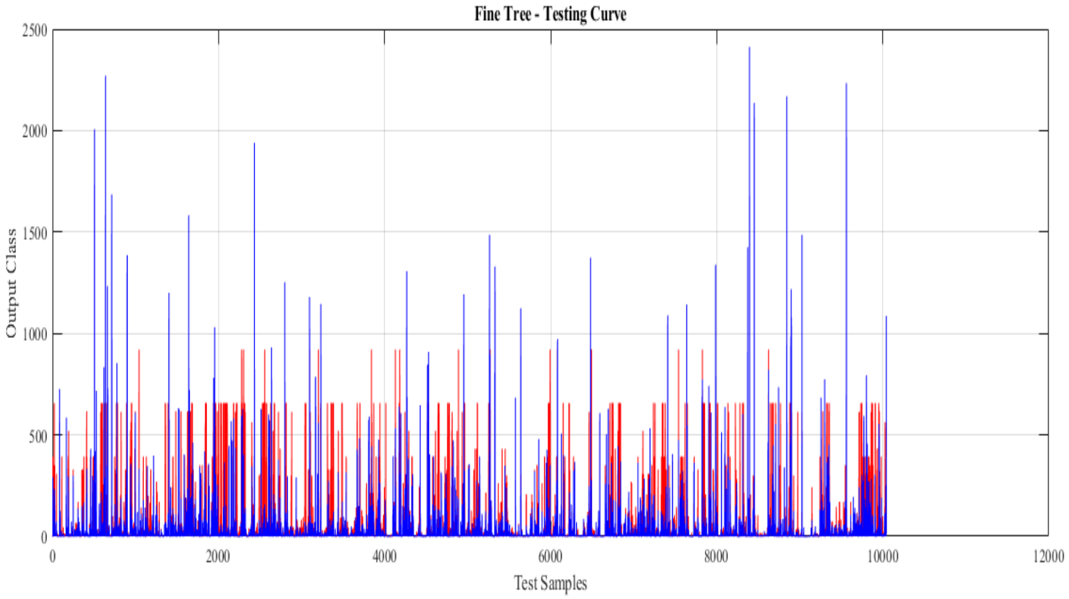

Fig. 1 Fine Tree

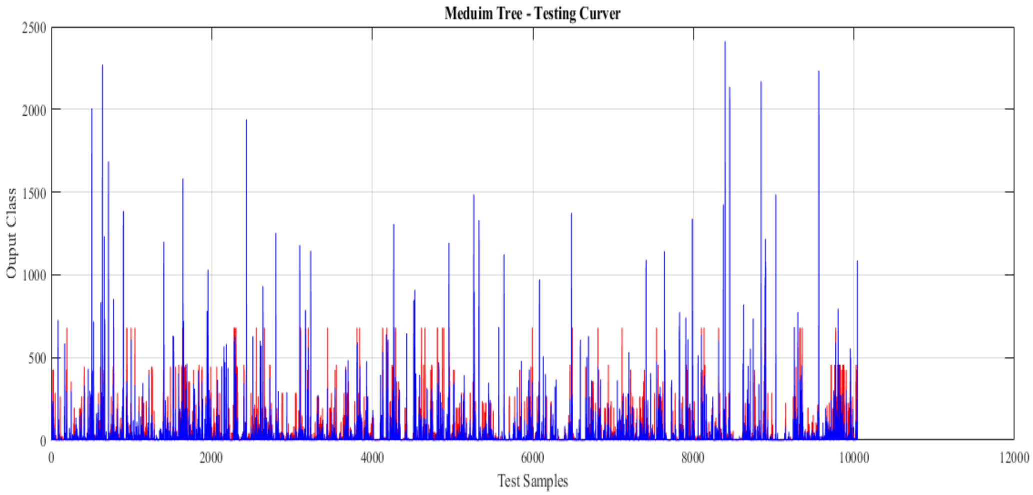

Fig. 2 Medium Tree Testing Graph

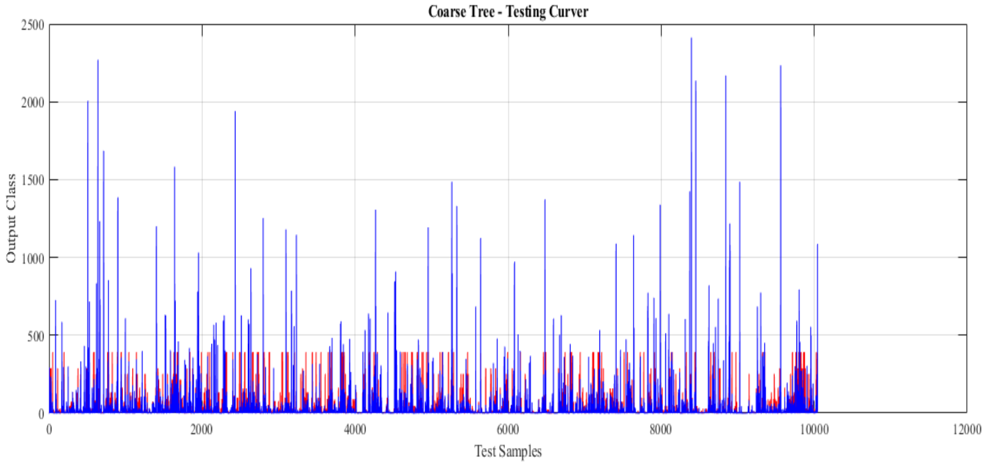

Fig. 3 Coarse Tree

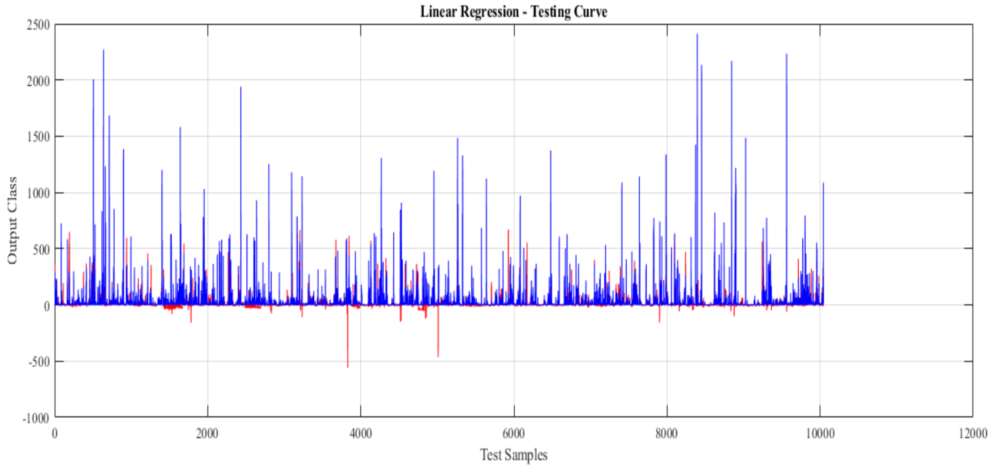

Fig. 4 Linear Regression Testing Curve

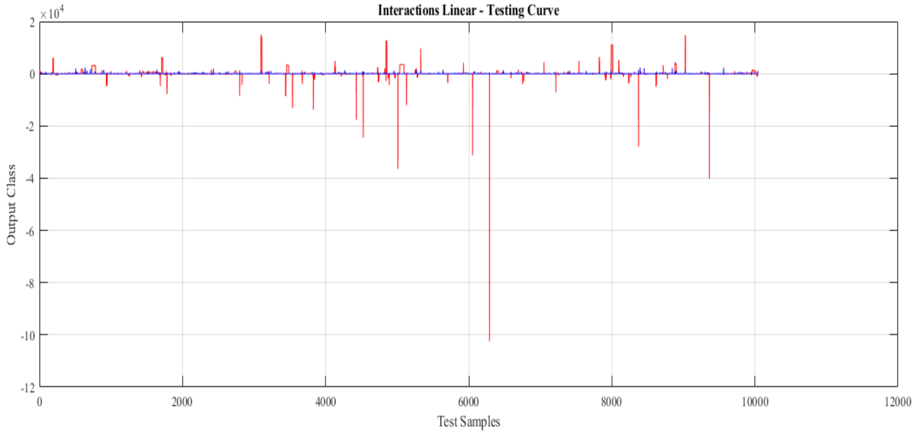

Fig. 5 Interactions Linear Testing Curve

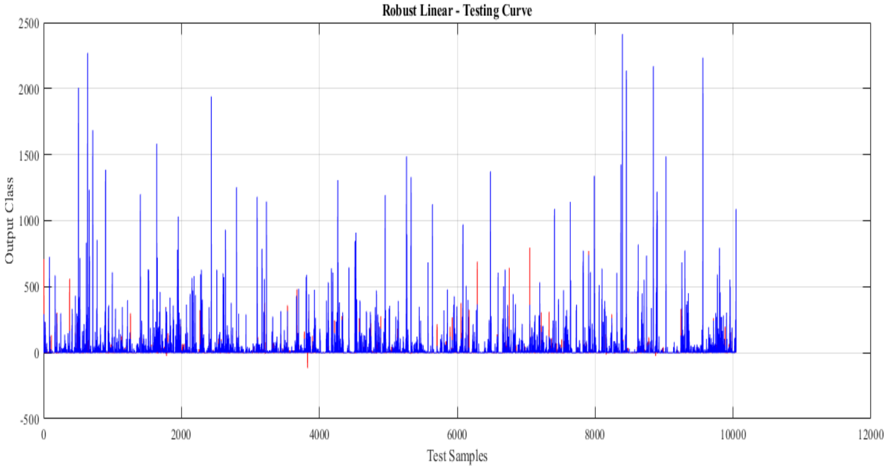

Fig. 6 Robust Linear Testing Curve

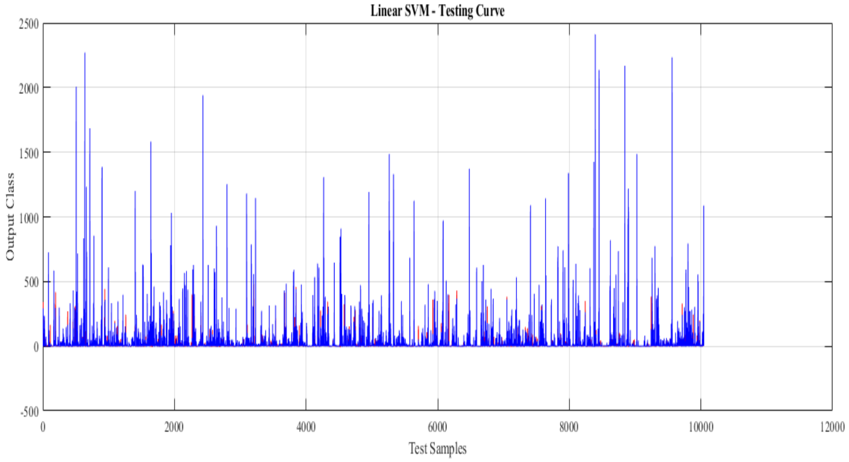

Fig. 7 Linear SVM

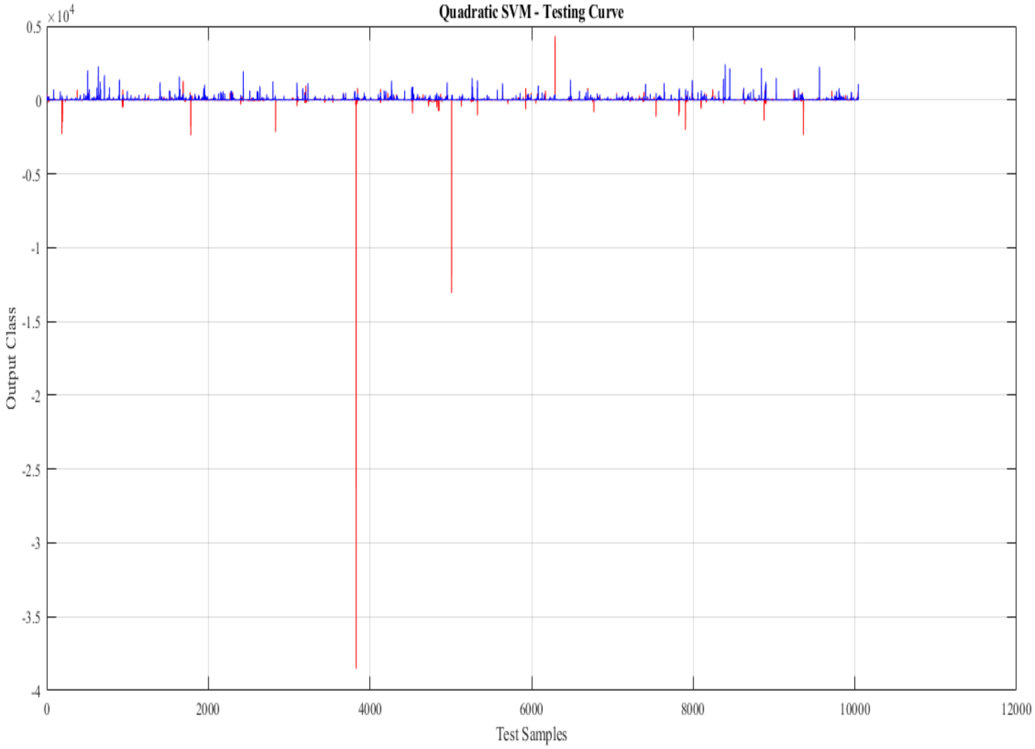

Fig. 8 Quadratic SVM Testing Curve

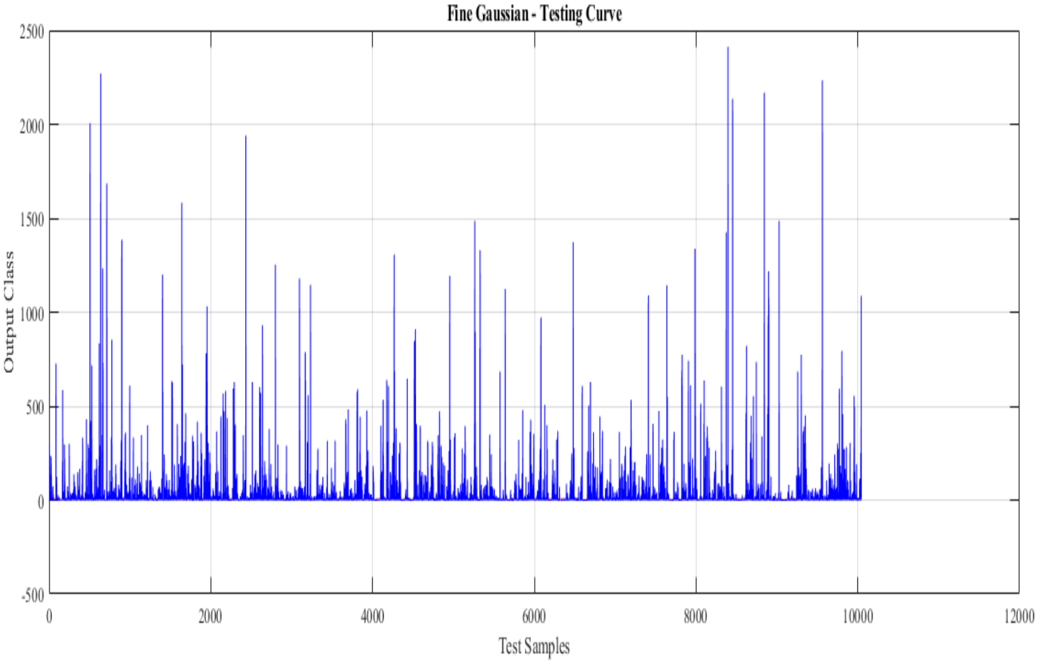

Fig. 9 Fine Gaussian Testing Curve

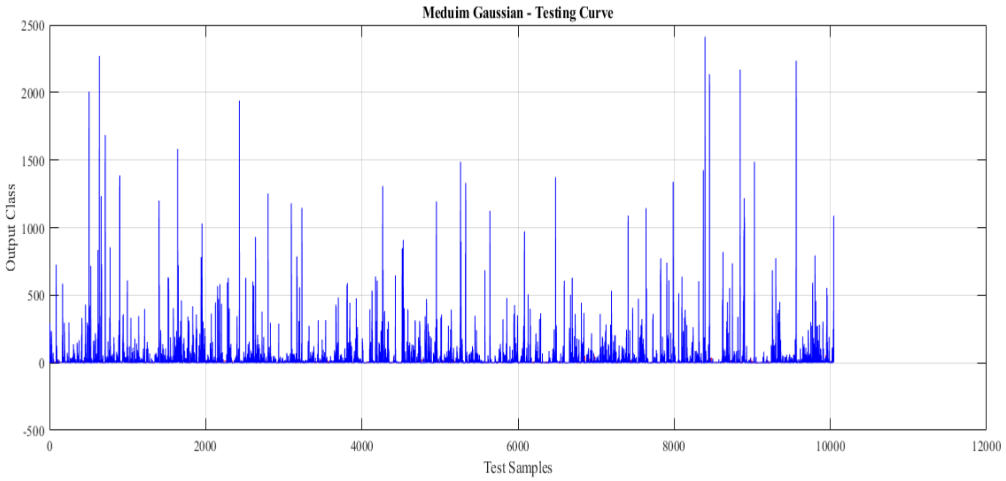

Fig. 10 Medium Gaussian SVM Testing Curve

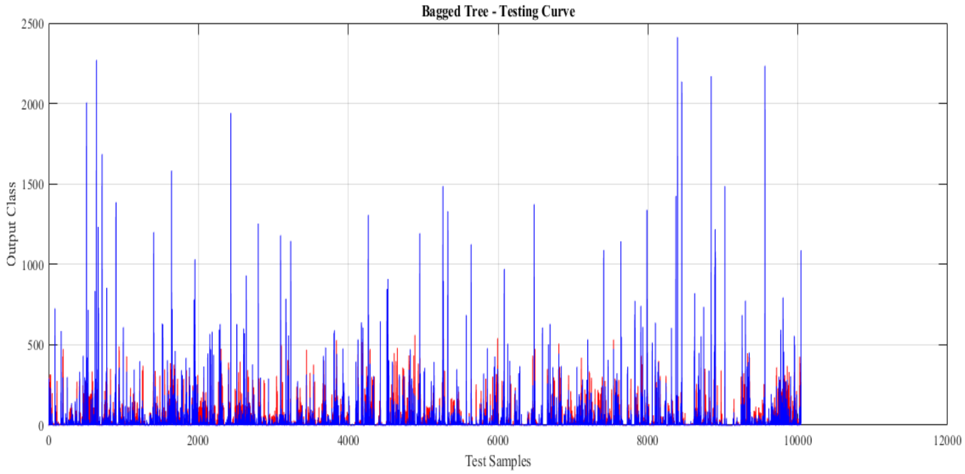

Fig. 11 Bagged Tree Testing Curve

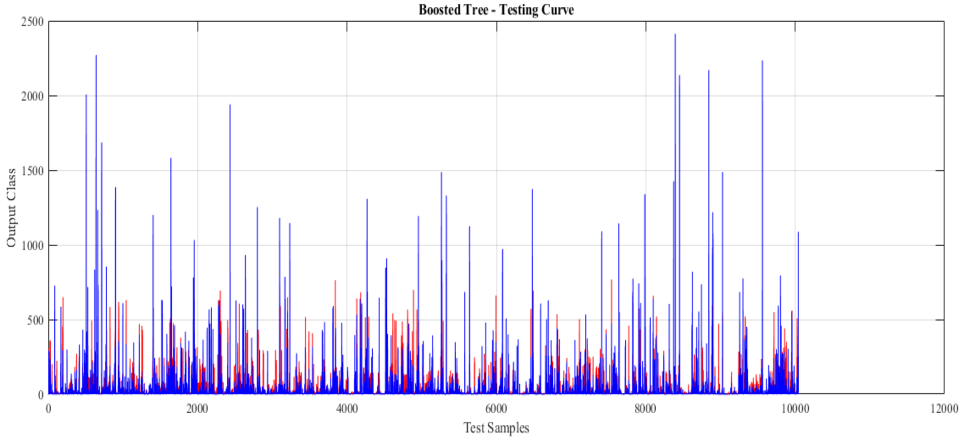

Fig. 12 Boosted Tree Testing Curve

V. RESULTS

After analyzing the results of all the algorithms, it was discovered that Fine Gaussian SVM had a testing accuracy of 100%, making it the best algorithm for the dataset.

However, after experimenting with various decision tree, SVM, ensembles of trees, and logistic regression algorithms, a new approach to mining data from social media platforms such as Facebook, Twitter, and Instagram has been discovered.

Figure 13 depicts a complete graphical representation of all the algorithms used, as well as the parameters used to evaluate them. Prediction speed varies significantly; algorithms like medium tree and coarse tree took the longest to predict, but had lower testing accuracy and higher errors.

On the other hand, while all SVM algorithms had lower errors and faster prediction speeds, only the fine Gaussian SVM was able to achieve 100% testing accuracy.

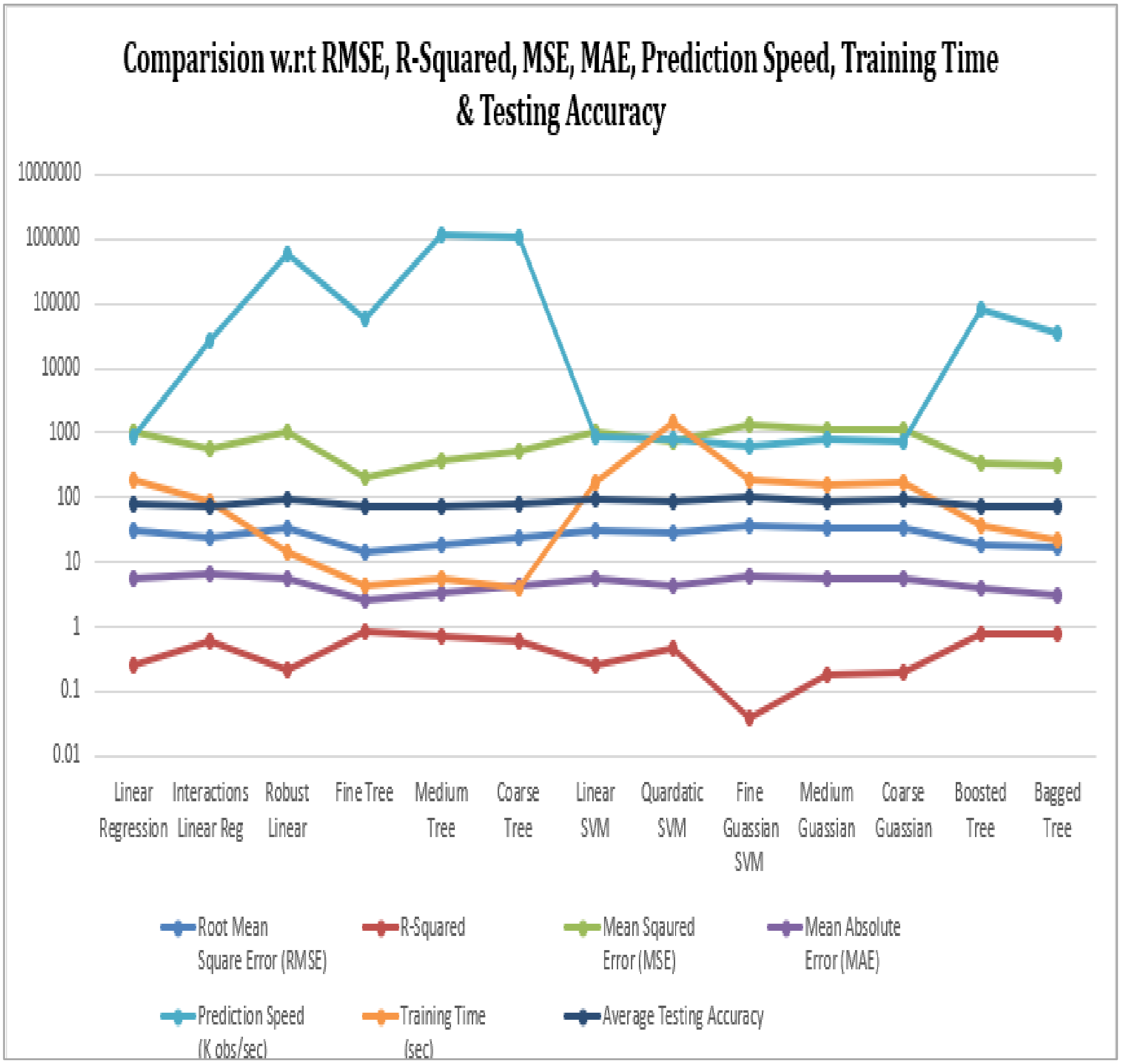

Fig. 13 Comparative Graph of Regression Algorithm

VI. CONCLUSION AND FUTURE WORK

Although researchers have previously looked at data from social media from a variety of angles. The dataset used in this study is the Facebook comment dataset, which includes 52 attributes that influence the number of comments a Facebook post is likely to receive. This paper is an attempt to find the best algorithm for predicting results in the said domain for such goals. Fine Gaussian SVM was found to give results that were equal to the actual results after applying thirteen different algorithms using MATLAB as a tool. When compared to other algorithms, the total testing accuracy was 100 percent with a lower percentage of errors.

REFERENCES

[1] R. Bhatia, "Here Are Facebook's Most Compelling Deep Learning Applications," June 2018. [Online]. Available: https://www.analyticsindiamag.com/here-are-facebooks-most-compelling-deep-learning-applications/.

[2] W. He, S. Zha and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," International Journal of Information Managemen, 2013.

[3] S. Forouzani, "Using social media and machine learning to predict financial performance of a company," 2016.

[4] D. Trottier and F. Chiristain, Towards a theoretical model of social media surveillance in contemporary society, 2015, pp. 113-135..

[5] A. Akay, A. Dragomir and B.-E. Erlandsson, "Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care," IEEE Journal of Biomedical and Health Informatics, 2015.

[6] T. Poell and J. V. Djick, Social Media and Activist Communication, 2015, pp. 527-537.

[7] K. Singh, R. Kaur and D. Kumar, "Comment Volume Prediction using Neural Networks and Decision Trees," in 17th UKSIM-AMSS International Conference on Modelling and Simulation, 2015.

[8] S. A. Bozkır, S. G. Mazman and E. A. Sezer, "Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case," IMCW 2010, CCIS 96, pp. 145-153, 2010.

[9] S. Abuzaid, Understanding Decision Trees – A, 2018.

[10] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. de Carvalho and A. A. Freitas, "A Survey of Evolutionary Algorithms for Decision Tree Induction," IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS - PART C: APPLICATIONS & REVIEWS, vol. X, no. XX, 2012.

[11] G. Mountrakis, J. Im and C. Ogole, "Support vector machines in remote sensing: A review," ISPRS Journal of Photogrammetry and Remote Sensing, Vols. Department of Environmental Resources Engineering, SUNY College of Environmental Science and Forestry, pp. 1-13, 1 November 2010.

[12] E. Ikonomovska, J. Gama and S. Džeroski, "Online tree-based ensembles and option trees for regression on evolving data streams," Neurocomputing, 2015.

[13] "Most popular mobile social networking apps in the United States as of July 2018, by reach," [Online]. Available: https://www.statista.com/statistics/579334/most-popular-us-social-networking-apps-ranked-by-reach/. [Accessed October 2018].

[14] S. A. Bozkir, S. G. Mazman and E. A. Sezer, "Identification of User Patterns in Social Networks by Data Mining Techniques : Facebook Case," Springer, pp. 145-153, 2010.

[15] Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, 2010.