Analysis of Social Network Comments Using Feature Selection and Random Forest Algorithms

Hamid H. Hussien [1†] , Mubarak H.Elhafian[1†] and Ahmed Hamza Osman [2††]

[1†] Department of Mathematics, College of Science and Arts, King Abdulaziz University, P.O. Box 344, 21911, Jeddha Rabigh, Saudi Arabia

[2††] Faculty of Computing and Information Technology, King Abdulaziz University, P.O. Box 344, 21911, Jeddha Rabigh, Saudi Arabia

hhahmed1@kau.edu.sa; mhelhafian@kau.edu.sa; ahoahmad@kau.edu.sa

Summary
Over the last decade and a half, online advertising on social networking sites have received considerable media interest. Data are being posted to these social networking sites every day. The highly dynamic behavior of users in relation to these services is therefore very important to study. In Facebook posts, user comments play a significant role in making decisions about which service or commodity are worth time and money. Due to many user comments being uploaded to these social networking services every day, and the growing value of these comments. This paper aims to analyze and predict user comment volume generated on Facebook prior to publication on the Facebook platform. We model the feedback from users and estimate how many responses a post will get over the next hour. We established a model prediction using the feature selection algorithm and the random forest model. In this situation, we consider the comments from short textual messages that refer to the main topic of the post. Our predictive model was used on numerous data sets, and the following parameters were measured: correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error estimation measurements. In mean absolute error criteria, our proposed methodology was more successful than the existing prediction models for Facebook comments with 24.40% rate.

1. Introduction

In recent years, the internet has made a significant progress and has immensely changed our economic and social life. Social networks (social networking sites or, simply, social media) are a successful example. Facebook, YouTube, Instagram, Qzone, Weibo, Twitter, Reddit, Pinterest, Ask.fm, Tumblr, Flickr, Google+, and LinkedIn are popular social networks that have attracted and connected millions of people all over the world. These online social networks enable users to construct a public or semi-public profile, articulate a list of other users with whom they share a connection, or view their list of connections and those made by others within the system[1]. They also allow individuals to freely generate and consume a huge amount of data such as uploading text, images, and videos to their profiles; comment on products; communicate their issues including health problems; and share many subjects or links with other users online. In the second quarter of 2019, based on data by Statista corporation, Facebook was the biggest social network worldwide, boasting about 2.41 billion monthly active users, and Twitter was among the top five social media platforms[2], boasting about 330 million monthly active users; it enables users to writes posts of up to 280 characters. Facebook, Twitter, and YouTube are the most widely used by companies to promote themselves and their brands; however, users are more likely to be connected through Facebook followed by YouTube, WhatsApp, and WeChat. Figure 1 shows the most popular social networks in July 2019, ranked by number of active users (in millions).

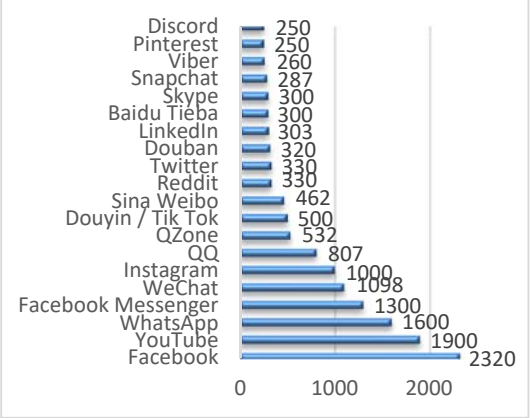

Fig. 1 Global ranking of social networks activities during July 2019

As mentioned above, social networks enable users to communicate with other users and share interests, political views, and information. This information can be utilized in marketing to improve a company's reputation and survival and increase sales revenue, in health care to improve service delivery, in politics to predict the results of public opinion poll, and in many other fields of social life[3]. In the last 3 years, the time people spend on Facebook has increased considerably as it became the biggest social networks. It had almost 2.5 billion monthly active users in the fourth quarter of 2019. Facebook provides fast interaction with many other people, resources that may increase or change the maintenance of interpersonal relations, and news[4].

Manuscript received April 5, 2020
Manuscript revised April 20, 2020

Facebook users can share interest and political views as well as write comments on other posts, such as news feed posts. Given the flood of social network data, there is considerable interest and opportunity for studying social networking comments.

This paper evaluates a machine learning–based approach in modeling Facebook user comments. Several options are available in the literature for predicting the comment volume of social network users, which usually model the pattern of the user comment over the posts in the past and predict the number of comments that the posts will receive over the next few hours. In this paper, we focus on analysis and model the global pattern of social network comments and create the most efficient and accurate model to predict the comment volume using feature selection and random forest (RF) algorithms. On social networking sites, commenting in the form of short textual messages is the method to discuss the topic of a post [5]. They are emerging and tightly connecting web users globally. Social network comments have attracted considerable interest of academic and companies intrigued by their communication web channel to entice their customers. Much of this interest lies in their importance in designing marketing and advertising campaigns as they play a key role in determining consumers' purchasing decisions[5]. User comments on various social network platforms can affect or change the perceptions of other users about the discussion topic or make the topic popular. For example, positive comments motivate people to update Facebook relationship status, whereas negative comments prevent them from updating Facebook relationship status. This means the observant's attitudes towards a relationship are more affected by the number of comments than by the actual nature of status[6]. User comments are considered important in software development by engineers to improve software quality. Apple's application stores consider user comments for every release[7]. User comments have been incredibly popular in term of online experience in recent years. They affect the perceptions of online content, as explained by Turner's theory, which states that users classify themselves and others by identifying the relevant group traits inside or outside the group when the shared social identity is psychologically prominent[8]. Social networks have become a major medium for social interaction. They provide features that may change the public discourse in society and set trends and agendas in topics such as environment, business and technology, and entertainment industry. These features have a significant impact on people's behavior in terms of communication and purchasing[9]. In addition, user comments also can helps companies understand the customer needs better. Thus, studying user comments on social media can create an improved research experience in field of opining maiming.

In this study, we aim to predict user comment volume generated by online news articles prior to publication of the post on Facebook. Following the standard knowledge discovery process (data collection, data preprocessing, feature construction, model training, and model testing), feature selection modeling and RF algorithms are used to build a model to estimate the expected comments that a post received in the next "n" hours. The performance of the models is evaluated using statistical metrics to choose the best model. The rest of this paper is organized as follows: Section 2 reviews works on comment volume prediction. Section 3 presents the mechanisms of the proposed model for predicting social network comments with evaluation function. The experimental design and datasets are reported in Section 4. Section 5 explains the results and discussion of this research. Paper conclusion and future work is demonstrated in Section 6.

2. Related Work

Many studies have evaluated the comment volume field using different social networks platforms and different statistical methods. Using raw Facebook data, Mandeep and Verma applied a linear regression and nonlinear regression hybrid model to predict the likelihood of the comment volume, i.e., the number of negative, positive, or neutral comments, that a post may receive in next hours [9]. The model set up included a data processor, a crawler, and an information revelation module. Their new hybrid model was an integration of two models: linear regression (PACE regression) and nonlinear regression (REP Tree)[9]. Their model performed better than existing methods in improving the time and space complexity along with accuracy using only significant features with low misclassification rate[9]. Singh et al. (2015) handled Facebook user comments using neural networks and decision trees to build up a forecasting model for the comment volume[10] and evaluated various dataset variables. They found that the accuracy of the decision tree model was higher than that of the neural network in predicting some measure of comment volume in a new blog post. The research was investigated the content of political blogs using linear regression, naive Bayes, elastic regression, and Topic-Poisson models and analyzed the relationship between the content and comment volume based on precision, recall, and F1 measure[11]. They focused on forecasting which blog post will get greater than the normal volume of response from users, measured in comments or words in comments. Their results suggest that the modeling topics can improve recall while predicting high-volume posts. Paul et al examined the influence of user comments on social media on people's opinion in the context of relationship status updates. They conducted an online experiment involving participants from Facebook users. Data were analyzed using a between-subjects 2 x 2 factorial design. They found that comments from other users alter opinions of a Facebook relationship status update[5]. Manos et al. predicted the user comments volume generated by

online news articles prior to publication time. The study was classified into two groups: generating comments, and receiving few or many comments[12]. The study suggested that models used comments in form of textual and semantic features have better and strong performers. In addition, the combination of all features leads to better and robust classification. Mishne and Rijke build a model to find a relationship between the actual mood and the mood of the blogosphere during given time intervals[13] based on textual features and temporal metadata of blog posts. Their models revealed a significant relationship between the two; also, the moods reported by the bloggers significantly improved compared with the baseline. Another study analyzed the content of online news agencies to examine the factors affecting the distribution of contents to public. Articles were classified into in three groups: (1) articles without comments, (2) articles with moderate comments, and (3) articles with many comments. Their model made predictions with >70% accuracy; the publish date and weight set up for content measure were the most informative features. This investigation can be generalized to other events on important days such as elections and more geographical features.

3. Research Methodology Framework

Many statistical methods have been used to predict the social network user comments; usually, they model the pattern of the user comments to past posts or documents and predict the number of comments that the post or document will receive in the next few hours. The design of the proposed method in the present study is illustrated in figure 1.

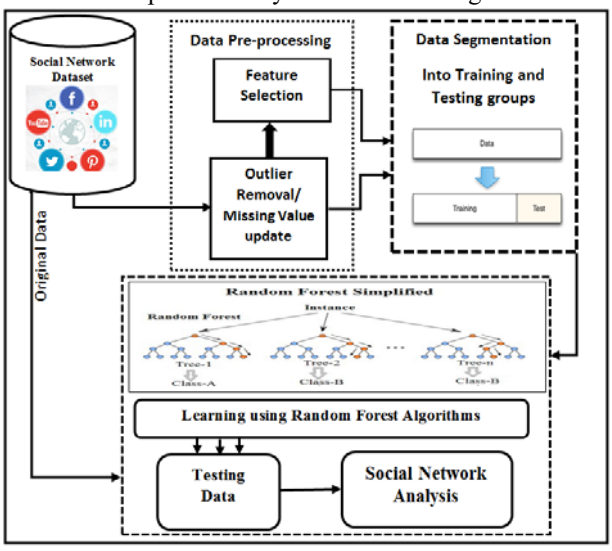

Fig. 2 Research Methodology Framework

In our model, the first stage is preprocessing—data cleaning, which includes outlier removal and updating missing values, and feature selection algorithm, which is used to select the important dataset features that can affect prediction. In the learning stage, the dataset is split into training and test data based on 10-fold cross validation to examine all the dataset. The RF algorithm was used in the final stage to test and predict the dataset. The proposed model aims to analyze and predict social media user comments to extract the best recommendation among the comments.

A. Feature selection

Feature selection is a machine-learning concept, which considerably affects the performance of the prediction model. It includes selection of a subset of relevant features for model construction, reduction of training times, simplification of the models, and improvement of the probability of generalization. It is usually used in the case of high-dimensional input information to minimize dimensionality in which the choice of the right set of functions for data modulation enhances supervised and unsupervised learning efficiency and reduces computational costs, such as learning time and resources required. In addition to reducing data size, it enhances the analytical phase of the data set by reducing the time to evaluate a large data set and eliminate noise from the data set[14]. The feature selection algorithm examines all combinations of Facebook comments from the dataset and introduces important features that contribute most to creating an effective classification system. This algorithm makes Facebook comments different from the original Facebook comment volume dataset. The feature selection method was first implemented to reduce the number of attributes[14].

B. The RF model

The RF model is a powerful machine-learning method for classification and regression in data mining; it was proposed by Breiman as a combination of decision trees[15]. Its core premise is to classify an element or an instance to a predefined set of classes based on their attribute values to reduce the error of the prediction[16]. The RF procedure is defined as follows: Let a collection of trees classifiers Φ_1 , Φ_2 ,…, Φ_i be independent and identical random vectors; each tree provides unit vote for the most popular class at input x . Mathematically, the equation is as follows:

$$\{f(x, \Phi_i), i = 1, \ldots N\} \quad (1)$$

where x is an input vector.

Equation (1) uses the bootstrap method to generate the maximum number of trees from the original training data set[17]. Because of its ability to handle a large number of data sets to model outcome predictions with good performance in reaching classification accuracy, RF models have been used in numerous studies on cyberbullying detection [18].

Manuscript received April 5, 2020
Manuscript revised April 20, 2020

4. Experimental Design and Dataset

It is feasible to acquire some generic training sets for the classifiers by using the methodology outlined in Section 3. This stage is completed before the final datasets were selected. In this paper, we use the Facebook comment volume datasets developed by Kamaljot et al. and reported in the UCI repository site[10]. The training data collection is a parallel method of variant comparisons and vectorization, and we gather three training sets by training collection preprocessing. The amount samples in these datasets are 40,949, 81,312, and 121,098 for sample Variant-1, sample Variant-2, and sample Variant -3, respectively. Each training variant samples is examined with 10 different test samples, and each group of tested samples is associated with different training variants. To estimate the feedback, user comment patterns are modeled on past posts. A model is trained, and predictions of how many comments a posts will receive in the next "n" hours are made. Data preprocessing is done by removing outliers and updating the missing information over 3 days from the chosen base moment or message without any remarks.

One of the main contributions of this study is to select the significant features affecting the prediction results. We used the feature selection algorithm to choose the most important features. The algorithm also allowed better computational process than using all the features of the dataset. The obtained results prove that the feature selection algorithm improved the prediction accuracy rate of our proposed model. The training and test phases were applied before and after selecting the most important features. The before-feature-selection accuracy was low compared with that after feature selection, which produced the degree of training depending on the number of important features extracted, and the reduction in the number of irrelevant features led to the increase in the correlation coefficient. The correlation coefficient rating was measured, and the Facebook comment volume dataset was used for training and testing. We have selected sometime in the past to predict feedback and reproduce as if the current time would be the selected time. We use the time of selection as the "base time." We know how many feedbacks the post received in the hours after the base time; that is, we know the estimations for the goal of these cases. As we understand the event after base time. At the same time, we only consider posts uploaded in the last 3 days as far as the base time is concerned because older pages do not usually have new comments. The regression analysis deals with this prediction problem. We used the RF algorithm in our prediction model before and after feature selection. Using this method, we predicted the number of feedbacks for the test data, with an aim to forecast the value of the target post.

5. Results and Discussion

Our experiments used two types of Facebook comment volume data (original and selected features). The original dataset is the typical comments data used in comments filtering, whereas the feature selection data set is created by using the feature selection algorithm within the original dataset. Selected data were used to test and differentiate between the patterns of the Facebook comments for each function. Therefore, only comments filtering can use the selected voting features based on the feature selection method. By selecting the important features, and thus reducing the number of features, the Facebook comments correlation increased. Different types of empirical studies focused on RF, and feature selection algorithms have been used to identify and Facebooks comments. The findings generated behind this hypothesis will be presented in a variety of stages: RF model with all features and RF model based on important features. The results section also presents the results of the cross-validation experimentation, correlation coefficient result, mean absolute error, root mean squared error, relative absolute error, and root relative squared error. The total correctly classified features were normalized by the total number of features. The classifying of the user comments has been computed using the following evaluation metrics.

- Pearson Correlation Coefficient (p_r)

With a range from 1 to -1, this metric computes the strength and direction of a linear relationship between two quantitative variables(x, y). The formula is

$$p_r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2]\,.[n\sum y^2 - (\sum y)^2]}} \quad (2)$$

where n is the number of data pairs.

- Mean absolute error (MAE)

MAE is the average absolute value of the mean, which can be calculated by the following formula:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right| \quad (3)$$

where $\hat{\theta}_i$ is the predicted value and θ_i is the actual value of the input variable.

- *Root mean squared error (RMSE)*

The RMSE is the square root of the variance that measures the performance of the model, and it is considered an excellent metric for numerical predictions. For N values of observation, the RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2} \quad (4)$$

where $\hat{\theta}_i$ is the predicted value and θ_i is actual target values of a variable.

- *Relative absolute error (RAE)*

RAE measures the performance of a predictive model. It is primarily used in machine learning, data mining, and operations management. It is expressed as a ratio:

$$RAE = \frac{\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{N}|\bar{\theta} - \theta_i|} \quad (5)$$

where $\hat{\theta}_i$ is the predicted values, θ_i is actual target values of a variable, and $\bar{\theta}$ is given by the following formula:

$$\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta_i \quad (6)$$

- *Root relative squared error (*RRSE)

RRSE is the average of the actual values. It is the total squared error divided by the total squared error of the predictor. Mathematically,

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^{N}(\bar{\theta} - \theta_i)^2}} \quad (7)$$

where $\hat{\theta}_i$, θ_i and $\bar{\theta}$ are as defined in equations (4) and (5).
The results are reported in Tables 1, 2, and 3 for datasets 1, 2, and 3, respectively. The results are presented with different training and test samples. The first results consist of 40,949 training samples with 10 test sample groups. These samples were classified and predicated based on the training prediction model. The average results for training and testing were calculated and reported as final. The average correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error achieved were 0.74409, 22.62918, 54.45944, 95.06913, and 77.72495, respectively. The correlation coefficient indicates the prediction results of the proposed model, and the mean absolute error indicates the mis-prediction of the proposed method with dataset. In the training stage, the correlation coefficient was 0.9798 and the mean absolute error was 1.8089. In the testing stage with sample number 7, we obtained a high correlation coefficient of 0.8725 and a low mean absolute error of 18.3416.

Table 1: Predication results of Variant Dataset No-1

Dataset No-1	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Training Dataset 1	**0.9798**	**1.8089**	**8.3826**	**16.3983**	**23.6167**
Test 1	0.7838	4.2211	**22.1562**	**38.2643**	62.4196
Test 2	0.8131	25.9735	53.4839	81.4387	61.1271
Test 3	0.8503	23.1541	47.2883	91.9232	54.0163
Test 4	0.7818	24.5634	49.3312	139.1939	106.5283
Test 5	0.5355	30.5339	113.9392	80.8056	82.7222
Test 6	0.79	22.1847	47.2976	80.1748	72.6484
Test 7	**0.8725**	**18.3416**	34.7113	123.3772	85.3634
Test 8	0.681	22.4655	52.3727	100.7675	74.1291
Test 9	0.483	25.6887	59.7492	137.8735	125.9872
Test 10	0.8499	29.1653	64.2648	76.8726	**52.3079**
Average	**0.74409**	**22.62918**	**54.45944**	**95.06913**	**77.72495**

Table 2 demonstrates the experimental results with another 81,312 training samples. These samples were inspected against 10 different samples in the testing stage. The average correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error were 0.70891, 26.85932, 71.58204, 102.3131, and 82.83295, respectively. In the training stage, the correlation coefficient
was 0.9809 and the mean absolute error was 1.7001. Moreover,
the highest correlation coefficient was obtained with test sample 10 (0.9247) and the lowest mean absolute error with test sample 6 (16.9206).

Table 2: Predication results of Variant Dataset No-2

Dataset No-2	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Training Dataset 2	**0.9809**	**1.7001**	**8.1206**	**15.7151**	**22.5264**
Test 1	0.7319	27.6824	62.5939	86.9124	71.5087
Test 2	0.817	25.1299	51.5001	100.009	58.8107
Test 3	0.7162	24.2606	56.8153	137.7878	122.614
Test 4	0.4795	33.0813	118.6749	87.6391	86.1426
Test 5	0.761	25.5969	51.5119	92.6304	79.0716
Test 6	0.8085	**16.9206**	32.3866	114.1845	79.607
Test 7	0.7289	22.2753	51.3916	100.2359	72.7146
Test 8	0.6047	46.3044	177.2932	113.7207	80.3442
Test 9	0.5167	23.709	65.7336	127.6471	138.522
Test 10	**0.9247**	23.6328	47.9193	62.3643	38.9933
Average	0.70891	26.85932	71.58204	102.3131	82.83295

Manuscript received April 5, 2020
Manuscript revised April 20, 2020

Table 3 illustrates the experimental results with another 121,098 training samples. These samples were inspected against 10 different samples in the testing stage consequently. The average correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error were 0.72877, 24.90468, 69.65356, 95.54436, and 81.8342, respectively. In the learning stage, the correlation coefficient was 0.9828 and the mean absolute error was 1.5459. Moreover, the high correlation coefficient the highest correlation coefficient was obtained with test sample 10 (0.9101) and the lowest mean absolute error was obtained with test sample 6 (15.5918).

Table 3: Predication results of Variant Dataset No-3

Dataset No-3	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Training Dataset 3	**0.9828**	**1.5459**	**6.9856**	**14.4879**	**20.872**
Test 1	0.7722	24.9632	58.5817	78.4435	66.9064
Test 2	0.8463	21.2816	47.9012	84.829	54.6908
Test 3	0.6753	24.2196	61.7761	137.7593	133.2667
Test 4	0.6777	30.6404	99.6021	81.2287	72.2883
Test 5	0.73	23.4139	53.0789	84.8053	81.443
Test 6	0.8174	**15.5918**	31.4571	105.4404	77.2967
Test 7	0.7228	20.2584	49.6566	91.353	70.2432
Test 8	0.6995	41.4457	174.9409	101.8881	79.2736
Test 9	0.4364	23.5863	66.2672	127.2488	139.5904
Test 10	**0.9101**	23.6459	53.2738	62.4475	43.3429
Average	0.72877	24.90468	69.65356	95.54436	81.8342

During each test, the data element cross tested one of the related training datasets. For all 10 parts, the overall results are calculated and reported for each variants dataset separately. The average correlation coefficient and mean absolute error were not too different between the three variant datasets: average correlation coefficient, 0.74409, 0.70891, and 0.72877, respectively, and average mean absolute error, 22.62918, 26.85932, and 24.90468, respectively. To confirm, we used analysis of variance (ANOVA) to compare the difference between these datasets (Variant Datasets 1, 2, and 3). We found no significant differences among the five evaluation metrics used in this study for the three datasets (all $P > 0.05$; Table 4).

Table 4: Comparison between the Facebook comments variant datasets using analysis of variance (ANOVA) test

Evaluation metrics	Source of variation	Sum of Squares	df	Mean Square	F	P-value
Root relative squared error	Between Groups	146.584	2	73.292	0.094	0.911
	Within Groups	21128.976	27	782.555		
	Total	21275.560	29			
Relative absolute error	Between Groups	328.391	2	164.196	0.248	0.782
	Within Groups	17880.968	27	662.258		
	Total	18209.359	29			
Root mean squared error	Between Groups	1759.212	2	879.606	0.636	0.537
	Within Groups	37318.846	27	1382.179		
	Total	39078.058	29			
Mean absolute error	Between Groups	89.642	2	44.821	0.812	0.454
	Within Groups	1490.164	27	55.191		
	Total	1579.806	29			
Correlation coefficient	Between Groups	0.006	2	0.003	0.173	0.842
	Within Groups	0.485	27	0.018		
	Total	0.491	29			

These results indicate agreement between the different datasets when classified by our proposed method. Figure 3 displays a comparison of the mean absolute error values between our proposed method using feature selection and RF algorithm and four other common prediction models.

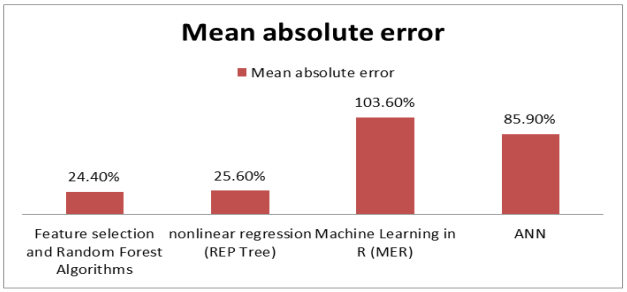

Fig. 3 Prediction comparison between feature-selected RF algorithm and other classifiers

Our proposed model achieved a 24.40% error rate, which is less than those obtained using other classifier techniques

such as neural network, MER in R, and nonlinear regression tree REP method[9].

6. Conclusions and future work

This research provides a classification scheme for social network comments based on Facebook posts from users. The project used three modules: (i) the Facebook users' list, (ii) preprocessing and feature reduction, and (iii) classification of user comments using RF algorithms. Feature collection has the advantages of raising the dimensions of the feature and reducing the running time. To predict the numebr of user responses on Facebook, the proposed model was classified using RF algorithms. The findings reveal that our model gave satisfactory results in terms of correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error. The suggested approach demonstrates that agreement between all variables used in the learning and evaluation processes by applying a significant ANOVA test. Nonetheless, the method has other constraints such as lack of an automatic mechanism to parse, clean, and store the contents of the comments and the inability to consider the visual and social meaning for more stable outcomes. In our future work, we will try to integrate an optimized approach to improve the prediction results and reduce the above limitations.

Acknowledgements

This work was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. G: 1510-662-1440. The authors, therefore, acknowledge with thanks DSR technical and financial support.

References

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," Journal of computer-mediated Communication, vol. 13, pp. 210-230, 2007.
[2] D. Noyes, "The top 20 valuable Facebook statistics," Zephoria, Florida, Available from: at https://zephoria. Com/social-media/top-15-valuable-facebookstatistics/[Accessed 10 February 2015], 2015.
[3] M. Bryant and J. Marmo, "The rules of Facebook friendship: A two-stage examination of interaction rules in close, casual, and acquaintance friendships," Journal of social and personal relationships, vol. 29, pp. 1013-1035, 2012.
[4] Tong and J. B. Walther, "Relational maintenance and CMC," Computer-mediated communication in personal relationships, vol. 53, pp. 98-118, 2011.
[5] P. W. Ballantine, Y. Lin, and E. Veer, "The influence of user comments on perceptions of Facebook relationship status updates," Computers in Human Behavior, vol. 49, pp. 50-55, 2015.
[6] L. V. G. Carreño and K. Winbladh, "Analysis of user comments: an approach for software requirements evolution," in 2013 35th International Conference on Software Engineering (ICSE), 2013, pp. 582-591.
[7] A. Hogg and S. A. Reid, "Social identity, self-categorization, and the communication of group norms," Communication theory, vol. 16, pp. 7-30, 2006.
[8] S. Asur and B. A. Huberman, "Predicting the future with social media," in 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, 2010, pp. 492-499.
[9] Kaur and P. Verma, "Fusion of PACE regression and decision tree for comment volume prediction," International Journal of Database Theory and Application, vol. 9, pp. 71-82, 2016.
[10] 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), 2015.
[11] T. Yano and N. A. Smith, "What's worthy of comment? Content and comment volume in political blogs," in Fourth international AAAI conference on weblogs and social media, 2010.
[12] M. Tsagkias, W. Weerkamp, and M. De Rijke, "Predicting the volume of comments on online news stories," in Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 1765-1768.
[13] G. Mishne and M. De Rijke, "Capturing Global Mood Levels using Blog Posts," in AAAI spring symposium: computational approaches to analyzing weblogs, 2006, pp. 145-152.
[14] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning," Computers & Security, p. 101710, 2020.
[15] Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.
[16] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," Expert Systems with Applications, 2019.
[17] J. C.-W. Chan and D. Paelinckx, "Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," Remote Sensing of Environment, vol. 112, pp. 2999-3011, 2008.
[18] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Detecting aggressors and bullies on Twitter," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 767-768.

Manuscript received April 5, 2020
Manuscript revised April 20, 2020

Associate Prof. Dr. Hamid H. Hussien. He acquired his B.Sc. degree in Econometrics and Social Statistics from the University of Khartoum. His M.Sc. in Statistics from Universiti Sains Malaysia, Malaysia and PhD in Statistics from Sudan University of Science and Technology, Sudan. He is a former dean of the faculty of commercial studies, Kordofan University, Sudan and a head of the Department of statistic, College of Science, University of Science and Technology. His researches interest includes: Modeling, Stochastic Processes, Multivariate Analysis, Applied Probability. Biostatistics, Safety research, Epidemiology. Currently, he is working in King Abdulaziz University (Rabigh Campus), College of Science & Arts, Dept. of Mathematics/ Rabigh, Saudi Arabia

Dr.Mubarak Hassan Mubarak Elhafian graduated with a Bachelor of Science(Applied Statistics) from Sudan University of Science and Technology , Department of Statistics, College of Science Sudan. He obtained his MSc Degree in Statistics from Sudan University of Science and Technology, Sudan. His received PhD degree in Statistics from Sudan University of Science and Technology, Sudan. He is currently work as Assistant Professor in King Abdulaziz University, Department of Mathematics, College of Science and Arts, Jeddah, Kingdom of Saudi Arabia. His current research interest includes Multivariate Analysis, Time Series Analysis, Stochastic Processes.

Assoc. Prof. Dr. Ahmed Hamza Osman graduated with a Bachelor of Computer Science from International University of Africa. He obtained his Master's Degree in Computer Science from Sudan University of Science and Technology, Sudan and his PhD in Computer Science with excellent academic achievements from Universiti Teknologi Malaysia (UTM). He was the Head of Computer Science department at the Faculty of Computer Studies at international University of Africa. Currently he works as Associate Professor in King Abduaziz University (KAU) Saudi Arabia. His research interest includes Information Retrieval, Plagiarism Detection, Soft Computing, and Data Mining, Natural Language Processing and Text Summarization.