Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 49-62
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/IJNAA.2021.4796

Predicting the number of comments on Facebook posts using an ensemble regression model

Omid Rahmani Seryasat[a*], Isaac Kor[b], Hossein Ghayoumi Zadeh[c], Arash shams Taleghani[d]

[a] *Assistant Professor, Department of Electrical Engineering, Shams Higher Education Institute, Iran.*
[b] *Department of Computer Engineering, Faculty of Technical Engineering, Shams institute of Higher Education, Gonbad Kavous, Iran.*
[c] *Assistant Professor of Biomedical Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.*
[d] *Assistant Professor, Aerospace Research Institute, Ministry of Science, Research and Technology, Iran.*

Abstract

The nature and importance of user's comments in various social media systems play an important role in creating or changing people's perceptions of certain topics or popularizing them. It has now an important place in various fields, including education, sales, prediction, and so on. In this paper, Facebook social network has been considered as a case study. The purpose of this study is to predict the volume of Facebook users' comments on the published content called post. Therefore, the existing problem is classified as a regression problem.

In the method presented in this paper, three regression models called elastic network, M5P model, and radial basis function regression model are combined and an ensemble model is made to predict the volume of comments. In order to combine these base models, a strategy called stack generalization is used, based on which the output of the base models is provided to a linear regression model as new features. This linear regression model combines the outputs of the 3 base models and determines the final output of the system.

To evaluate the performance of the proposed model, a database of the UCI dataset, which has 5 training sets and 10 test sets, has been used. Each test set in this database has 100 records. In the present study, the efficiency of the base models and the proposed ensemble model is evaluated on all these sets. Finally, it is concluded that the use of the ensemble model can reduce the average correlation coefficient (as one of the evaluation criteria of the model) to 74.4 ± 16.4, which is an acceptable result.

Keywords: regression, ensemble model, Facebook, and comment volume prediction

*Corresponding Author: Omid Rahmani Seryasat
Email address: orseryasat@shamsgonbad.ac.ir (Omid Rahmani Seryasat[a*], Isaac Kor[b], Hossein Ghayoumi Zadeh[c], Arash shams Taleghani[d])

Received: 30 June 2020 *Revised:* 5 October 2020

1. Introduction

Machine learning methods have a wide range of applications [1-14]. Social networks are a new generation of sites that are the focus of users of global Internet networks these days. Such databases are based on online organizations and each brings together a group of Internet users with a specific feature. Social networks are considered to be a type of social media that has made it possible to achieve a new form of communication and content sharing on the Internet. Social networks can be explored and evaluated from different aspects. In this article, an attempt is made to predict comments on Facebook posts.

The strong presence of social networks in cyberspace as one of the most basic and most widely used web pages, has caused a large number of users to spend part of their daily life in cyberspace. Although there have been various conversations and statements in various fields related to social networks, especially Facebook, the complexity of the phenomenon of second life in cyberspace, part of which takes place in social networks, is a bed of all kinds of new life. This has caused the circle of conversations about these networks to be very large and diverse. Social networks, led by Facebook, are one of the new phenomena of cyberspace and the new life of millions of people with the largest number of users. So far, nearly one billion people have joined Facebook.

The user of cyberspace has no limits to communicate on Facebook. This network in the strict sense of the word has become a virtual country, where residents from all over the world are accepted, entry and exit to this virtual country does not require a visa and every human being can enter this country only by having an Internet connection. It is a country where real people are present in cyberspace and sometimes with their avatars.

Public social networks, such as Facebook, provide users with a variety of features depending on the purpose of their activity, such as sharing text, images and multimedia files, adding friends, and the system of sending and receiving messages. Facebook allows the user to deactivate this part of the system if he does not want to receive messages. Other features of this network include easy connection by all personal electronic devices, such as mobile phones, tablets, etc. This makes it easy for the users to log in to their personal page anywhere in the world and entertain themselves by adding posts or reading other people's posts and similar activities. As a result, people are increasingly turning to social networks such as Facebook.

In the social networking service, the most active and important of them, namely "Facebook pages" are selected for analysis. It is intended to determine the number and type of comments of a post in a few hours and the number of comments it will receive. Studying the behavior and opinions of social network users can provide very useful information on various fields, such as cultural, political, tourism and even e-commerce and marketing issues.

The overall purpose of this study is to provide a model to predict and identify the volume of comments of social network users on a post in the next few hours. In this research, the most active social network service, Facebook, is studied and analyzed. The main purpose of this research is to analyze the tendencies and patterns of Facebook users.

In this regard, this general goal is divided into several sub-goals:

- Identifying appropriate features for behavioral analysis and modeling of user preferences
- Determining the efficiency of these forecasts in different areas
- Determining the appropriate model of user behavior and tendencies
- Investigating the impact of users' comments on each other's opinion

Finally, by examining and comparing the classification methods, it is tried to reveal their strengths and weaknesses, and as a result, analyze the appropriate methods to cover the weaknesses and improve the strengths.

The volume of comments on social media can be measured as the number of words in the comments section, the number of comments, the number of distinct users who have submitted comments, or various other types. These actions can be influenced by various factors, such as the original text of the document/post, links to other documents/posts, time of day a post appeared, side conversation, page likes, page check, or page category, and etc.

2. Literature review

For comment volume prediction, patterns of user comments appear on the posts/documents, and forecasts are presented based on the number of comments. A limited number of research has been done on predicting comment volume using different social media platforms, which will be briefly described here.

Different regression models, such as reduced error pruning, m5p tree, perceptron multilayer, and RBF network can be used to predict comment volume. Singh et al. (2015) worked on Facebook using neural networks and decision trees and provided a software example that shows the volume of comments. These evaluations are performed on different types of data, so the decision tree performs better than neural networks in predicting the volume of descriptions. Similarly, Buza (2014) demonstrates an industrial-conceptual evidence that automatically analyzes documents in Hungarian blogs. In this article, the author uses different features of blogs to teach different regressor models and evaluates the results using the criteria of @ 10Hits and AUC @ 10. The result shows that the regression model is better than simple models.

Even classifiers can be used to categorize comments in specific classes, such as the study by Tsagkias et al. (2009). This article reports on predicting the volume of comments on news articles before publication using a random forest classifier based on a set of five features of surface, cumulative, textual, semantic, and real-world features. They do this in two steps. First, the binary classification of articles with the potential to receive comments, and second, the classification of articles into ”low volume” and ”high volume”. Outputs indicate better binary classification results and evaluate that textual and semantic features are stronger than other features. Similarly, the study of Balali et al. (2013) analyzed the content and timing of online news agencies to identify factors influencing the dissemination of content to the public. They also used the random forest classifier to categorize articles into three categories of without comment, intermediate comments (1-6) and very high comments (¿ 6). The proposed model accurately predicts more than 70% and reports that the release date and weight introduced for content measurement were more informative than other features. Results can be predicted based on important days (i.e. elections, festivals, and holidays) and geographical features. While the study of Tsagkias et al. (2010) shows the dynamics of user opinions in seven different news sites, this article discusses the two-factor distribution of log normal and negative functions and predicts the volume of comments using a linear model and active comparison throughout the various news sites. The results show that the prediction of the volume of long-term comments may be with a small error after 10 source-hours in the observations.

Jamali et al. (2009) worked on Digg.com social bookmarking website. They defined collaboration and participation of users of a network and studied the behavioral characteristics of users using the comment information. Moreover, they measured the entropy and concluded that users on the Digg.com site were interested in a wide range of topics. Using a classification and regression framework, the popularity of online content was predicted based on theoretical data and social network-derived features. This article reports a 1 to 4 percent loss of classification accuracy. This is despite the fact that the forecast is based on the popularity metric and uses only a few hours of comment data compared to all available comment data. The results can be further improved by polar analysis of comments.

Different thematic models can be used to extract hidden topics in post content. Yano et al. (2010) studied political blogs using a variable thematic model and analyzed the relationship between content and comment volume. The Naive Boys model is also used for binary forecasting, i.e. high volume or low volume, and the forecast is evaluated according to precision measurement, recall, and F-criteria. They concluded that predicting high-volume posts could improve learning topics.

Even, Negi et al. (2012) predicted the state of user-content links of Flickr groups to show the chance of one user commenting or another user updating an image. They considered both the social effect using the Transactional Mixed Membership Stochastic Block (TMMSB) and the content effect using the Latent Dirichlet Allocation (LDA) to predict user-content links. Time zone effects can be used in the future to obtain more accurate results.

The most popular social networking site, Facebook, is used in the study of Rahman (2012), who offers a data mining architecture for collecting social data. This article collects various features, such as personal information, comments, wall posts, and age using the Facebook API key. It then compares information about age groups for different uses as predicting human behavior, job responsibility distribution, pattern recognition, decision making and product promotion. The nearest neighbor k algorithm is used to classify numerical and textual properties and range properties of values such as number of posts on the wall, number of ages, number of music, and number of interests at different levels of the class.

Summarizing users' comments is more difficult when mixed with different opinions. Especially in the case of restaurants where the overall rating of the restaurant is evaluated based on different comments about different foods. Zheng et al. (2015) has proposed a new approach to summarize the comments on restaurants. They used real-world reviews to purchase from the most popular Chinese and English restaurant websites, Dian-ping and Yelp. Using food features and user comments, as two independent dimensions in the hidden space, it has created a two-way thematic model combined with word-extraction algorithms related to opinion expression and clustering-based selection algorithms. Their method provides a high quality summary of restaurant dishes. This concept can be used for broader applications, such as selling different goods or services.

3. The proposed method

In this section, the method suggested in this article is presented to predict the number of comments related to each Facebook post. In this regard, a teachable ensemble model was used in which 3 regression models called MP5 tree model, radial basis function and elastic model were combined. In the proposed ensemble model, a simple linear regression model is used to combine the output of each of the base models. In addition, in this paper, we will use an effective feature selection method based on correlation to remove irrelevant and redundant features. The details of the proposed methods are described below.

3.1. M5p tree model

A decision tree model called M5 was proposed to predict continuous data. This model, unlike other common decision tree models that present discrete classes as output, creates a multivariate linear model for the data in each node of the tree model. The structure of decision tree models includes the steps of creating a tree and pruning it. In the tree construction stage, an inferential algorithm or division criterion is used to generate a decision tree. The division criterion for the M5 model algorithm is the evaluation of the standard deviation of the values of a class that arrives at a node as a quantity of error and calculates the expected reduction in this error as a result of testing each feature in that node. The reduction of the standard deviation is obtained from the following relation:

$$SDR = sd(T) - \Sigma \frac{|T_i|}{|T|} sd(T_i) \quad (1)$$

Where, T represents the series of samples that reach the node, Ti represents the samples that have the i^{th} output of the potential series, and sd represents the standard deviation. Due to the branching process, the data in the child nodes have less standard deviation than the mother node and are therefore purer. After maximizing all possible branches, $M5$ selects an feature that maximizes the expected reduction. This division mostly forms a large quasi-tree structure that causes overfitting. To overcome the problem of overfitting, the formed tree must be pruned. This is done by replacing a subtree with a leaf. Therefore, the second step in designing a tree model involves pruning the grown tree and replacing the subtrees with linear regression functions. This tree model production technique divides the input parameters' space into smaller areas or subspaces and fits a linear regression model in each of them. After the linear model is obtained, the model is simplified to minimize the estimation error by removing the parameters. The M5 uses a greedy search to remove variables that have little involvement in the model. Of course, sometimes all variables are removed and only one constant value remains. Figure 1 shows how the M5 decision tree model works for a hypothetical problem. Each model represents a linear regression equation. For example, if X1¿2.5 and X2¿2, then the third model is used in the form of $Y = a_0 + a_1X_1 + a_2X_2$.

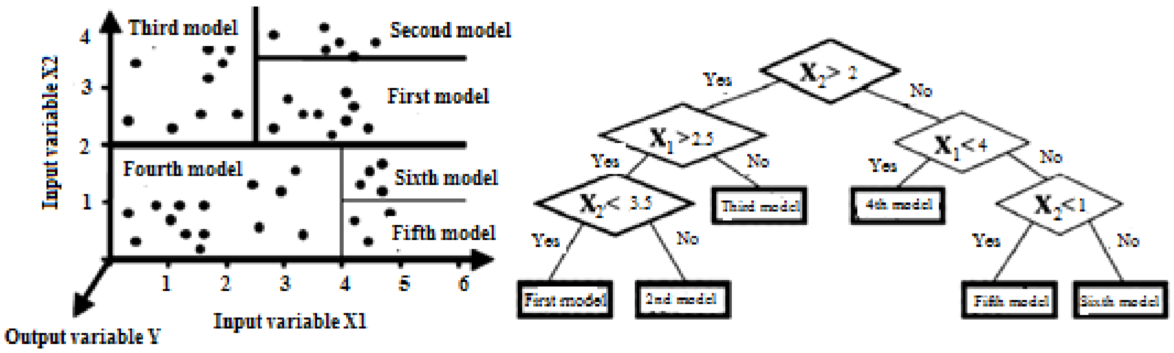

Figure 1: $m5$ model performance for decision tree construction. Left figure: Division of the $x1$ and $x2$ parameters' space into 6 areas. Right figure: expression of the criterion of dividing the space of input parameters into a tree

3.2. Supervised Gaussian radial basis function

Radial basis function (RBF) networks are a type of feed forward neural network with a long history. However, relatively few articles have covered how to train such a network so that it can make accurate predictions. A common strategy is to train the hidden layer of the network using the k-means clustering algorithm and to train the output layer of the network using supervised learning. But in this paper, it is showed that if the hidden layer of the network is trained using a supervised algorithm, more accurate predictions will be made. In this approach, learning center positions, local variances of basic functions, and feature weights are adjusted in a supervised manner. In the RBF predictor, the error squares along with a polynomial penalty on non-bias weights in the last layer of the network were used as a loss function to find the network parameters.

The Gaussian radial basis function model is defined as follows:

$$f(x_1, x_2, \dots, x_m) = g\left(w_0 + \Sigma_{i=1}^{b} w_i exp\left(-\Sigma_{j=1}^{m} \frac{a_j^2 (x_j - c_{ij})^2}{2\delta_{ij}^2}\right)\right) \quad (2)$$

Where x_1 to x_m are the feature values' vector for a given sample, g(.) is the activation function, b is the number of base functions, w_i is the corresponding weight for each base function, a_j^2 is the j^{th} feature weight, c_i is the center of the base function, and $\delta_{i,j}^2$ is variance functions. It should be noted that this model is the most difficult model to implement because both the weight of the features and the variance corresponding to each basic function for each feature must be determined during the training process. By selecting the appropriate parameter settings, simplified versions of the above model can be used. By default, feature weights are not used (i.e. a_j^2 is considered a constant value of 1) and there is only one global variance parameter (i.e. $\delta_{i,j}^2 = \delta_i^2$). Therefore, in this case, there is only b parameter of variance. Setting the parameters of $w(l)$,i, a_j^2, $c_{i,j}$ and $\delta_{i,j}^2$ is done after determining a local minimum on the squares of the training data error. This error function is calculated as follows:

$$L_{SSE} = \left(\frac{1}{2}\Sigma_{i=1}^{n}(y_i - f(\bar{x_1}))^2 + (\lambda \Sigma_{i=1}^{b} w_i^2)\right) \quad (3)$$

In the above relation, y_i is the target value for the training sample of x_i. As can be seen, the first series is performed on all n training samples. The λ parameter is known as the ridge parameter and its role is to control the share of penalties on the weights in such a way as to prevent the model overfitting. This parameter can be set by the user.

The gradients of the error functions are composed of corresponding partial derivatives in comparison to the network parameters. These derivatives are calculated using a standard account. Calculating partial derivatives, just like multilayer perceptron, involves post-fault propagation in the network. In the implementation, there are two gradient-based methods for optimizing the parameters: the Quasi-Newton method using BFGS and the nonlinear conjugate gradient reduction method.

Before the training phase begins, all numerical features are normalized and mapped to values between 0 and 1. Even the target feature (i.e. the output value) is subject to this normalization, but after determining the predicted value, the predicted normal number is displayed in the initial non-normal space. In the implementation of this regression method, unknown values of numerical features are replaced by the average value of that feature and unknown values of nominal features are replaced by the mode value. Constant features are removed and nominal features are converted to binary. The same steps are followed for the test sample, whose output should be predicted.

Another important step in network training is the initialization of network parameters. In the implementation, the initial values of network weights are randomly extracted from a uniform distribution in the range of 0.25 to -0.25. This sampling strategy is practically adapted from a well-known heuristic, according to which it is better to randomly weigh the weights with a small amount.

Initialization of hidden unit centers and variances is a much more complex process than initialization of network weights. Since the k-means clustering algorithm is a fast algorithm that is frequently used to train the hidden layer of the RBF network, this algorithm has been used to initialize the hidden layer centers ($c_{i,j}$). In addition, the initial value of all parameters of $\delta_{i,j}^2$ variance is considered equal to the maximum square of the Euclidean distance between the centers of the clusters. According to this approach, we will make sure that the initial value of the variance parameter is not too small. This approach, in practice, leads to a resilient learning process. It is also important to note that when using a_j^2 features' weights, these values are initialized.

In the implemented model, if any delta value is less than a user-defined threshold limit (10-6) when calculating gradients, then the delta value is considered zero and further calculations are avoided.

3.3. Elastic network

Another model used in this article is a model called elastic network. The motivation for creating this network stems from solving the problem of another model called Lasso. Suppose we show our dataset as (X, y) that X is the prediction matrix with dimensions $n \times p$ and$y$ is the output vector. In Lasso, the goal is to minimize the following objective function:

$$\underset{\beta}{min} \|y - X\beta\|^2 \quad s.t. \quad \|\beta\|_1 = \Sigma_{j=1}^{p} |\beta_j| \leq t \quad (4)$$

In the above relation, if$p > n$, the lasso model selects a maximum of n variables. That is, the number of selected features is limited by the number of samples. Another problem with Lasso is that it cannot perform a group selection. This means that it selects one variable from a group and ignores the rest. The elastic model tried to solve the problems of Lasso model using the following equation:

$$\hat{\beta} = \underset{\beta}{argmin} \|y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \quad (5)$$

The norm part of a penalty leads to the production of a sparse model. The squared part of a penalty has several effects, which are:

- Removing the limit on the number of selected variables
- Increasing the impact of feature grouping
- Stabilizing the first norm path

It should be noted that the path produced by the elastic network is a piecewise linear. The proposed objective function in the elastic network model is solved by a stepwise algorithm called LARS-EN and the final path is generated.

3.4. Combination of predictors

In this section, one of the most important innovations of this article, which is the design of a new ensemble predictor, is discussed. The design of the ensemble is basically based on a principle called Diversity. This means that the basic predictors used should be somehow diversified. This diversity can be done at four data, feature, predictor, and combiner levels. Diversity at the data level means that the basic predictors use different samples for training. Diversity at the feature level means that not all basic predictors are trained on a fixed set of features and use different feature sets for training. The third level is based on the principle that different predictors are used for training. In fact, predictors that are taught differently must be used. The last level is related to the combiner. For example, several types of strategies, such as majority voting, weighted voting, and the use of a learner can be combined, thus creating diversity. In this article, it is intended to use a new strategy to combine predictors.

In the present study, the stack generalization method is used to train the proposed predictor. The idea of stack generalization is as follows. Suppose Z is the name of the dataset that contains N samples with different outputs. In our case, these outputs are the number of comments per post. In the proposed method, we divide the Z dataset into 10 separate sets. We also assume that we use all three regression models introduced in the previous section. In the stack generalization method, we train each model using the standard 10-fold cross-validation method. Database X is randomly divided into 10 non-overlapping sections of equal size, X_i ,$i = 1, 2, \dots, 10$. To produce each pair of train and test data, one of the 10 sections is used for testing and the other 9 sections for training. This operation is repeated 10 times. In this way, 10 pairs are obtained as follows:

$$
\begin{array}{l}
V_1 = X_1 \quad , \quad T_1 = X_2 \cup X_3 \cup \cdots \cup X_{10} \\
V_2 = X_2 \quad , \quad T_2 = X_1 \cup X_3 \cup \cdots \cup X_{10} \\
. \\
. \\
. \\
. \\
V_{10} = X_{10} \quad , \quad T_{10} = X_1 \cup X_2 \cup \cdots \cup X_9 \quad (6)
\end{array}
$$

At the end of the training procedure, there are 10 copies of each predictor, each version being trained on 10 datasets.

Next, we need a separate predictor to learn how predictors vote. To this end, for each sample in the $X1$ subset, the outputs generating the trained predictors on the$T1$ subset are retained and used as new features to construct the desired dataset for the combiner. Thus, the three outputs generated from the base predictors (RBF, M5P, and ElasticNet) and the actual sample output in the $X1$ form a new feature vector. With these interpretations, the training dataset for the combiner predictor is a 4-element vector. In the following, the same process is repeated for the samples in the $X2$ subset, and using the output of the trained predictors on $T2$, a series of other 4-element vectors are added to the combiner training set. The same process is repeated for other subdivisions. Once the combiner training set is complete, we need to train our combiner on this dataset. In this paper, we used linear regression as a combiner.

After combiner training, the 10 subsets are re-integrated into the Z set, and the basic predictors are re-trained. In this way, both the basic predictors and the linear ensemble predictors are prepared to predict the output of the test samples. As can be seen in Figure 2, when a sample X is prepared for testing, the trained predictors provide their prediction of the sample X to the combiner. Thus, a true 3-element real vector is assigned to the ensemble predictor to determine the final output.

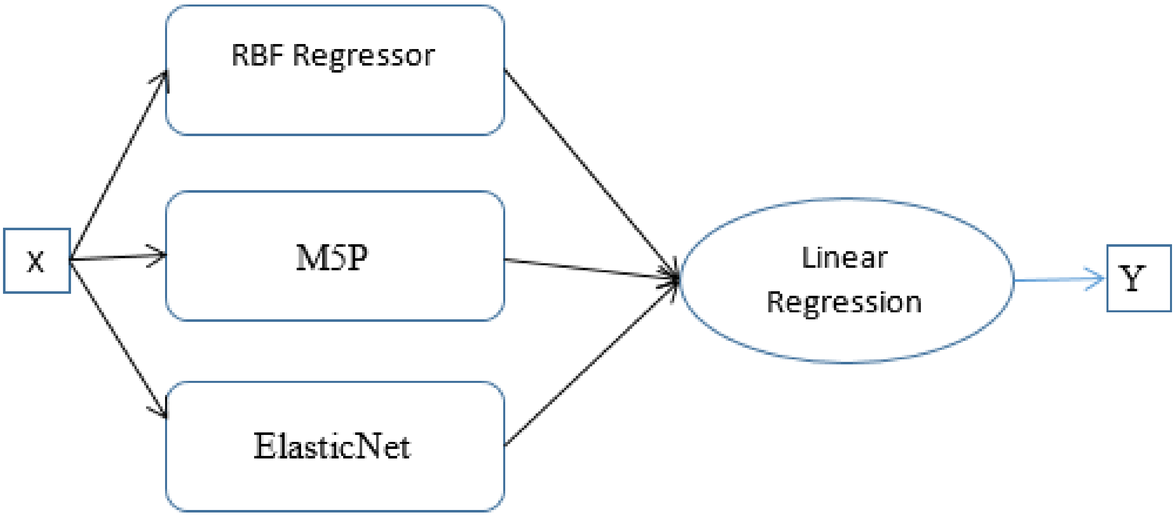

Figure 2: Architecture of the proposed ensemble predictor

3.5. Feature selection

In the database used in this paper, 52 features have been extracted for each sample. Now, it is attempted to reduce the number of these features and select a suitable subset of them.

The purpose of feature selection is to select the k features from among the existing d features that provide the most information. Thus, we omit the (d-k) features. In order to select the best subset, this article uses the correlation-based feature subset selection (CFS) criterion. According to this competency criterion, a subset is more appropriate whose members are highly correlated with the relevant output and, on the other hand, are not dependent on each other. According to this principle, irrelevant features should be removed because they have little correlation with the output of the samples. On the other hand, redundant features should be removed because they are highly correlated with one or more features and it is sufficient to use one feature instead of all of them. The equation for this competency criterion is as follows:

$$M_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \quad (7)$$

Where, M_s is a heuristic criterion of the competency of a feature set S that includes k features, $\bar{r}_{cf}$ is the mean output-feature correlation ($f \in S$), and $\bar{r}_{ff}$ is the mean feature-feature correlation. The fraction numerator indicates the ability of the features to predict the output and the denominator of the fraction also indicates the redundancy between the features. In the present study, a genetic algorithm has been used to search for a suitable subset.

In solving a problem using genetic algorithm, 2 basic steps of definition of chromosome and definition of fit function are proposed. The first step in starting a genetic algorithm is how to show a chromosome that indicates a candidate solution. We have a simple task ahead for the feature selection problem. Here, 62 features have been extracted, so the chromosomes will have 52 genes. If the value of a gene is equal to 0, it means that the corresponding feature of that gene has not been selected, and if it is equal to 1, it means that it has been selected. The criterion used in the CFS method is also used to define the fit function. The following figure shows the general block diagram of the genetic algorithm.

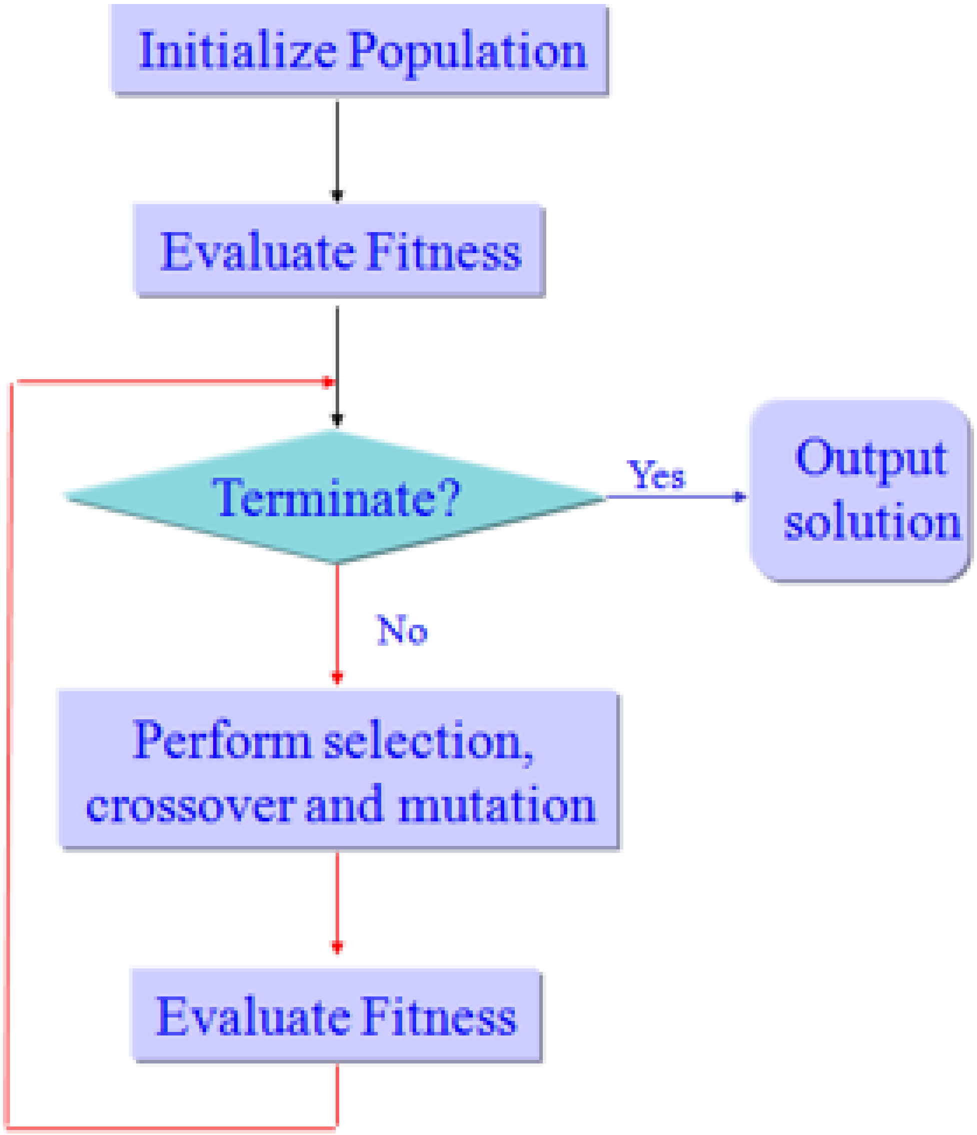

Figure 3: Block diagram of feature selection method

4. Results

In this section, we will first introduce the dataset used in this article. Then, we investigate the effect of the proposed ensemble algorithm on improving the accuracy of the final regression. Finally, the rate of improvement of the accuracy and speed of the proposed algorithm by the proposed feature selection method is evaluated.

4.1. Dataset

The number of samples in the dataset used in this article is 40949 samples. These samples were made in 2016 from the contents of Facebook pages. In this dataset, which is available online in the UCI data warehouse, 53 input features plus an output feature (i.e. the number of comments below that content) are extracted from each Facebook post (i.e. each sample). Table (1) summarizes the extracted features for each sample. As can be seen, these features can be divided into four groups.

The features of the first group try to measure the popularity and effectiveness of the source of an article. The second group of features, known as essential features, display the pattern of comments below each post at different time intervals. The third group is related to some features related to the document, such as: document length, time gap between the base time and the time the article was published, page promotion status, number of shares, and etc. The fourth group of features is known as features of the days of the week, in which the day post was published and the day of the week showing the base time are shown as two seven-element binary vectors.

Table 1: Description of the features in the dataset used in the article

No.	Group	Name	Description
1	First	Number of likes per Facebook page	This feature indicates the popularity of the page and the level of public support for a particular page.
2	First	Number of Facebook page check-in	This feature shows how many people have visited a particular section on Facebook so far.
3	First	Talk about page	This feature estimates people's daily interest in the source of an article / document. This feature is calculated based on the number of people who like a page and return to it. It actually determines the actual number of users of a page.
4	First	Page category	This feature specifies the type of document source: location, institute, brand, and so on.
5-29		Derivative features	These features are extracted by the page through the calculation of the minimum, maximum, average, median and standard deviation of the essential features.
30	Second	CC1	Total number of comments before selecting base time.
31	Second	CC2	Number of comments in the last 24 hours (after base time).
32	Second	CC3	Number of comments between the last 24 hours and the last 48 hours (after base time).
33	Second	CC4	Number of comments in the first 24 hours after the publication of the article, but after the base time.
34	Second	CC5	Difference between CC2 and CC3.
35	Third	Base time	This property is between 0 and 71 (hours).
36	Third	Content length	The number of characters in a post that indicates its size.

37	Third	Page sharing number	This feature shows how many people have shared the relevant content among their friends.
38	Third	Page promotion status	This feature is a binary feature that indicates whether the person has advertised their content or not? Because some people on Facebook advertise their content to get more views and pay for it.
39	Third	H-local	This feature defines the time at which the target comments are received.
40-46	Fourth	The day of the week when the post was published	These features are a seven-element binary vector that specifies what day of the week this article was published.
47-53	Fourth	Basic week day	This feature shows the day of the week as the base time. This feature vector contains seven binary elements.
54		Output	The number of comments in the H next hours (H is expressed in feature 39).

4.2. How to measure the efficiency of regression methods

The dataset in the UCI data warehouse divides the prepared samples into 5 training groups and 7 test groups (there are 100 samples in each test set). In this article, we will report the results of the relevant tests on all test sets for training on each of the 5 training sets. Therefore, in the studies performed, 50 tests are performed. In each test, the efficiency of the regression model was mentioned based on conventional criteria, such as correlation coefficient, mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), and root relative squared error (RRSE). In the following, the formula of the mentioned criteria is presented.

$$Correlation\ coefficient = \frac{\Sigma_{i=1}^{n}(a_i - \bar{a})(p_i - \bar{p})}{\sqrt{\Sigma_{i=1}^{n}(a_i - \bar{a})^2 . \Sigma_{i=1}^{n}(p_i - \bar{p})^2}} \quad (8)$$

$$MAE = \frac{\Sigma_{i=1}^{n}|p_i - a_i|}{n} \quad (9)$$

$$RMSE = \sqrt{\frac{(p_i - a_i)^2}{n}} \quad (10)$$

$$RAE = \frac{\Sigma_{i=1}^{n}|p_i - a_i|}{\Sigma_{i=1}^{n}|a_i - \bar{a}|} \quad (11)$$

$$RRSE = \frac{\Sigma_{i=1}^{n}(p_i - a_i)^2}{\Sigma_{i=1}^{n}(a_i - \bar{a})^2} \quad (12)$$

In the above equations, pi, ai, $\bar{a}$ and $\bar{p}$ show the prediction values for the i^{th} sample, the actual values for thei^{th} sample, the mean of the actual values, and the mean of the predicted values, respectively.

4.3. Prediction precision report

In this section, the results obtained from the methods used in this article are reported. It should be noted that the weka software was used to implement the mentioned methods. Normalization of features is essential in most machine learning techniques. In this article, the normalization of features has been done from the pre-processing part related to the software. In this regard, using unsupervised filtering of features, the values of all features are mapped to a value between zero and one. Next, irrelevant and redundant features are removed from the training set samples based on the genetic algorithm-based approach introduced in Chapter Four. In Table (2), the selected features from each training set are specified separately.

Two interesting points are taken from Table 2. The first point is that out of the 52 features extracted, a maximum of 7 features are extracted as useful features from the training sets, and more than 85% of the features are irrelevant or redundant that will cause the predictor performance to decline. The same result can be considered an interesting result. For example, the number of likes on a page is ineffective in predicting the number of comments. While everyone thinks that there is a strong correlation between the number of likes and the number of comments, but after applying the feature selection method, we came to the conclusion that this is not the case. The second point that can be deduced from this table is the common features selected in different training sets. For instance, features 30 (i.e. the total number of comments before selecting the base time), 33 (i.e. the number of comments in the first 24 hours after posting, but before base time), and 36 (i.e. the number of characters in the content indicating the post size) have been selected as effective features in all sets. The total number of different features selected for the 5 different training sets is only 10 features.

Table 2: Selected features in the various training sets available in the database

Training set	Number of selected features
First	11, 12, 30, 33, 36, and 49
Second	11, 12, 30, 33, 36, and 49
Third	16, 17, 26, 30, 33, 36, and 45
Fourth	12, 30, 33, 36, and 49
Fifth	12, 30, 33, 36, and 49

After selecting the feature, it is time to predict the number of comments based on the selected features. As stated in Section 3, three basic predictive models called RBF, M5P and Elastic-Net are combined in this paper. The results of each of these models and the result of their combination are reported in Table (3). As can be seen, the proposed ensemble model has performed better than other methods (especially based on an important criterion such as the correlation coefficient).

Table 3: Comparison of the precision of the proposed ensemble method compared to the basic methods used individually

Precision / Model	Correlation coefficient (%)	MAE	RMSE	RAE	RRSE
M5P	55.4±24.6	28.4±8.8	81.3±53.6	107.4±26.7	99.9±40.6
RBF	56±23.3	28.7±9.3	81.7±53.2	109.6±29	101.8±43
Elastic Net	55.5±25	27.7±8.8	81.2±53.9	106.8±26.8	99.8±41
Proposed model	74.4±16.4	27.6±8	78.6±32.1	105.5±36.7	99.5±44

5. Conclusion

An ensemble regression model was used to predict the volume of comments in this paper. To implement the proposed model, the UCI database was used, in which 53 features were extracted for each Facebook post. First, the available features were mapped to a range between zero and one using normalization. Then, a criterion called CFS along with genetic algorithm was used to remove irrelevant and redundant features. After the feature selection stage, it was observed that more than 80% of the features were irrelevant and redundant, the removal of which increases the efficiency of the proposed regression model. After selecting the useful features, an ensemble model, including 3 basic regression models (named elastic model, M5P and RBF), was proposed. In the proposed ensemble model, the output of these base models was provided to another regression model (in this paper a simple linear regression model) as new features. To create a new training set for this linear regression, a strategy called stack generalization was used. After training the proposed ensemble regression, this model was trained and tested on different sets. Based on the results, it is concluded that the combination of basic models using the proposed approach can lead to improved results. It is worth mentioning that the proposed model achieved an average correlation coefficient of 0.74, which is considered as an acceptable and promising result.

References

[1] K. Buza, L. Schmidt-Thieme, & R. Janning (Eds.), Feedback Prediction for Blogs. In M. Spiliopoulou, Data Analysis, Machine Learning and Knowledge Discovery (pp. 145-152). Cham: Springer International Publishing.(2014).

[2] M. Ghane, AR. Nejad, M. Blanke, Z. Gao, T. Moan, Statistical fault diagnosis of wind turbine drivetrain applied to a 5MW floating wind turbine, Journal of Physics: Conference Series 753 (5), (2017).

[3] M. Ghane, MJ. Tarokh, Multi-objective design of fuzzy logic controller in supply chain, Journal of Industrial Engineering International 8 (1), 1-8.

[4] M. Ghane, M. Zarvandi, MR. Yousefi, attenuating bullwhip effect using robust-intelligent controller, 2010 5th IEEE International Conference Intelligent Systems, 309-314.

[5] H.Ghayoumi Zadeh , A. Montazeri ,I. Abaspur Kazerouni , J. Haddadnia , Clustering and screening for breast cancer on thermal images using a combination of SOM and MLP. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2017 Jan 2;5(1):68-76.

[6] S. Jamali, H. Rangwala, Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis. Paper presented at the International Conference on Web Information Systems and Mining, (2009).

[7] E. Koozegar, M. Soryani, and I. Domingues. "A New Local Adaptive Mass Detection Algorithm in Mammograms." BIOSIGNALS. 2013.

[8] E. Kozegar, "Computer aided detection in automated 3-D breast ultrasound images: a survey." Artificial Intelligence Review (2019): 1-23.

[9] M. M. Rahman, Intellectual knowledge extraction from online social data. Paper presented at the Informatics, Electronics Vision (ICIEV), (2012).

[10] O. RahmaniSeryasat, J. Haddadnia. "Evaluation of a new ensemble learning framework for mass classification in mammograms." Clinical breast cancer 18.3 (2018): e407-e420.

[11] O. RahmaniSeryasat, J. Haddadnia, H. Ghayoumi-Zadeh, A new method to classify breast cancer tumors and their fractionation. Ciência e Natura, 37(4), (2015), 51-57.

[12] O. RahmaniSeryasat, J. Haddadnia, H. Ghayoumi Zadeh, Assessment of a Novel Computer Aided Mass Diagnosis System in Mammograms, Iranian Journal of Breast Disease 9 (3), (2016),31-41.

[13] O. RahmaniSeryasat, J. Haddadnia, "Assessment of a novel computer aided mass diagnosis system in mammograms." Biomedical Research (2017) Volume 28, Issue 7.

[14] A. Salmasi,A. Shadaram, A.S. Taleghani, Effect of plasma actuator placement on the airfoil efficiency at poststall angles of attack, IEEE Transactions on Plasma Science, 2013, 41(10), pp. 3079–3085, 6601652.

[15] S.M.Sheikholeslam Noori, M. Taeibi Rahni, S.A. Shams Taleghani, Multiple-relaxation time color-gradient lattice Boltzmann model for simulating contact angle in two-phase flows with high density ratio, European Physical Journal Plus, 2019, 134(8), 399.

[16] K. Singh, R. Kaur, &D. Kumar, Comment Volume Prediction Using Neural Networks and Decision Trees. Paper presented at the Proceedings of the 2015 17th UKSIM-AMSS International Conference on Modelling and Simulation, (2015).

[17] S. Negi, & S. Chaudhury, Predicting User-to-content Links in Flickr Groups. Paper presented at the Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), (2012).

[18] A.S. Taleghani, A. Shadaram, M. Mirzaei, S. Abdolahipour, Parametric study of a plasma actuator at unsteady actuation by measurements of the induced flow velocity for flow control, Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2018, 40(4), 173.

[19] M. Tsagkias,W. Weerkamp, & M. d. Rijke, Predicting the volume of comments on online news stories. Paper presented at the Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, (2009).

[20] M. Tsagkias, W. Weerkamp, &M. de Rijke, News Comments:Exploring, Modeling, and Online Prediction. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S, (2010).

[21] T. Yano ,NA. Smith, What's Worthy of Comment? Content and Comment Volume in Political Blogs. Paper presented at the Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, (2010).

[22] I. Zare, A. Ghafarpour, H. Ghayoumi Zadeh, J. Haddadnia, and M. Mostafavi Isfahani. "Evaluating the thermal imaging system in detecting certain types of breast tissue masses." (2016).

[23] R. Zhang ,Z. Zhang , X. He, & A. Zhou, Dish Comment Summarization Based on Bilateral Topic Analysis. Paper presented at the 2015 IEEE 31st International Conference on Data Engineering, (2015, April).