



REGRESSION METHODS FOR SOCIAL MEDIA DATA ANALYSIS


Dahiru TANKO, Department of Digital Forensics Engineering, Technology Faculty, Fırat University, Elazığ, Turkey
191144117@firat.edu.tr

( <https://orcid.org/https://orcid.org/0000-0001-7376-3306>)


Turker TUNCER, Department of Digital Forensics Engineering, Technology Faculty, Fırat University, Elazığ, Turkey
turkertuncer@firat.edu.tr

( <https://orcid.org/0000-0002-5126-6445>)

Sengul DOĞAN, Department of Digital Forensics Engineering, Technology Faculty, Fırat University, Elazığ, Turkey
sdogan@firat.edu.tr

( <https://orcid.org/0000-0001-9677-5684>)

Erhan AKBAL*, Department of Digital Forensics Engineering, Technology Faculty, Fırat University, Elazığ, Turkey
erhanakbal@firat.edu.tr

( <https://orcid.org/0000-0002-5257-7560>)

Received: 25.11.2021, Accepted: 09.05.2022

Research

*Corresponding author

DOI: 10.22531/muglajsci.1028299

Abstract

In the early 2000s, the more traditional modes of communication via mobile devices were voice calls, emails, and short message services (SMS). Nowadays, communication through mobile applications such as WhatsApp, Facebook, Twitter, Instagram, etc. About Facebook the leading social network with monthly active users of about 2.85 billion people. With this number of users, a large amount of data is generated. Exploring this data provides an insight into the users' activities which can aid in tackling security challenges and business planning, among other benefits. This study presents a neighborhood component analysis (NCA) and relief-based weight generation methods for a regression task on Facebook data. The features are calculated using the weight generated and four widely used activation functions. The features are then fed to four regression models for prediction. The proposed model is used to predict nine different attributes of the FB dataset whose values are continuous. RMSE, R-squared, MSE, MAE, and training time were calculated and used as evaluation metrics for all nine cases. The average R-square value of the Relief and NCA-based methods were calculated as 0.9689 and 0.9667, respectively. The results indicated that our proposed methods are very efficient and successful for regression tasks on Facebook data.

Keywords: Facebook, regression, neighborhood component analysis, machine learning

SOSYAL MEDYA VERİ ANALİZİ İÇİN REGRESYON YÖNTEMLERİ

Özet

2000'li yılların başında, mobil cihazlar aracılığıyla geleneksel iletişim yolları olan sesli aramalar, e-postalar ve kısa mesaj servisleri(SMS) kullanılırdı. Günümüzde WhatsApp, Facebook, Twitter, Instagram vb. mobil uygulamalar aracılığıyla iletişim tercih edilmektedir. Facebook, yaklaşık 2,85 milyar kişinin aylık aktif kullanıcısıyla önde gelen sosyal ağ uygulamasıdır. Yüksek kullanıcı sayısı ile büyük miktarda veri üretilir. Bu verileri keşfetmek, diğer faydaların yanı sıra güvenlik sorunlarının ve iş planlamasının üstesinden gelmeye yardımcı olabilecek kullanıcıların faaliyetlerine ilişkin bir fikir sağlar. Bu çalışmada, Facebook verileri üzerinde regresyon tabanlı bir komşu bileşen analizi (NCA) ve rahatlamaya dayalı ağırlık oluşturma yöntemleri sunulmuştur. Özellikler, oluşturulan ağırlık ve yaygın olarak kullanılan dört etkinleştirme işlevi kullanılarak hesaplanmıştır. Özellikler daha sonra tahmin için dört regresyon modeline beslenir. Önerilen model, değerleri sürekli olan FB veri setinin dokuz farklı özniteliğini tahmin etmek için kullandık. RMSE, R-kare, MSE, MAE ve eğitim süresi hesaplanmış ve dokuz vakanın tümü için değerlendirme ölçütleri olarak kullanılmıştır. Relief ve NCA temelli yöntemlerin ortalama R-kare değeri sırasıyla 0.9689 ve 0.9667 olarak hesaplanmıştır. Sonuçlar, önerilen yöntemlerin Facebook verileri üzerindeki regresyon görevleri için çok verimli ve başarılı olduğunu göstermiştir.

Anahtar Kelimeler: Facebook, regresyon, komşu bileşen analizi, makine öğrenmesi

Cite

Tanko, D., Tuncer, T., Dogan, S., Akbal, E., (2022). "Regression Methods for Social Media Data Analysis", *Mugla Journal of Science and Technology*, 8(1), 31-40.

1 Introduction

Nowadays, social media applications are widely used by all age group people [1-4]. These applications work on different platforms, namely smartphones, tablets, and personal computers with the internet [5-9]. On estimate, about 3.96 billion people are currently using social media. The most important reason for this widespread use worldwide is the people's desire to communicate independently. People can share their information, pictures, videos, and location with their friends in the virtual environment through different applications. Among these applications, the most widely used social media application at the moment is Facebook, with over 2.85 billion monthly active users as of July 2021 (Statista Research Department August 2, 2021). It was founded in 2004 by Mark Zuckerberg, while he was a student at Harvard University [10]. At the same time, the simplicity of the Facebook application interface is an essential factor for its phenomenal success. On Facebook, the list of friends means the people the person has included in the private domain in the virtual environment. In this area, people can share their personal information such as birthdays, school information, memories, and pictures [9, 11]. The figure below shows the social networks ranked by the number of active users in millions.

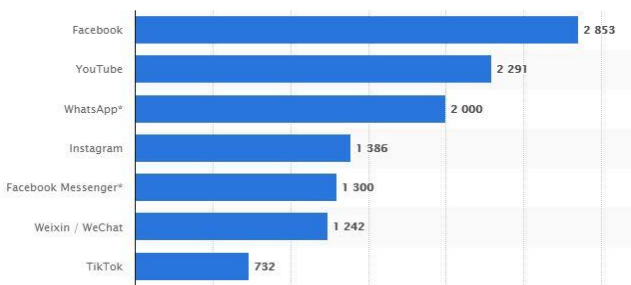


Figure 1. Global social networks ranked by number of active users (million)

Facebook has become a great network with links to friends. People can comment and add to particular topics. All these features significantly increased the amount of data in the virtual environment [12, 13]. The identification and analysis of relationships in this data are of great importance for interpreting people or situations. Analyzing the relationship between attributes of a Facebook dataset can be done through regression-based methods [14]. Regression is one of the widely used supervised learning methodologies for continuous data. It detects and reveals the relationship between many variables. So far, many regression models have been proposed and used to interpret many datasets and predict their labels [12, 15, 16]. Some of the state-of-the-art regression-based methods are given below to overview literature.

Ertugrul et al. [17] proposed a method using recurrent extreme learning machines and recurrent neural networks. Their system predicted the electricity load using root mean square error, R-square, and computation time as the evaluation parameters. The

proposed method showed high success in terms of speed and accuracy compared to the extreme learning machines. Delgado et al. [18] presented 77 regression methods, namely random forests, neural networks, Bayesian models for different datasets. They analyzed the results using Best-Squared Correlation, Root Mean Square Error (RMSE), and mean absolute error. The primary purpose of their study is to present a new method suitable for regression problems. The authors suggested some areas of application, the advantages, and the disadvantages of each method. [19] proposed a decision tree method using a non-linear regression method, and the success of this method is analyzed using squared errors, computational and space complexity. Erp et al. [20] explored linear regression methods for Bayesian penalization. In their study, six different conditions with regression coefficients. Median prediction, mean squared error, optimal credibility intervals, Matthews' correlation coefficient, correct and false inclusion were used to evaluate the results. Ertugrul and Tagluk [21] proposed a regression method termed as dependent nearest neighbor, which is similar to k nearest neighbor. The difference between the proposed method and the k nearest neighbor is the distance matrix. Four separate datasets were used for the evaluation of the model performance. They employed accuracy and computational costs to measure the performance of the proposed regression method. Prashanth [22] et al. proposed a prediction method to detect Parkinson's disease. They used logistic regression alongside support vector machine (SVM) models in their work.

In this work, we present a new feature extraction model for regression of social media data, with datasets acquired from Facebook. The characteristics of the proposed method are given below.

- ReliefF and NCA are used to generate weights.
- Four widely used activation functions and weight generated are used together to calculate the features.
- Four regression methods (linear, tree, support vector regression - SVR and Gaussian) are used to predict the output.
- Five evaluation metrics (RMSE, R-Squared, MSE, MAE) are used to measure the performance of the models

The significant contributions of the proposed method are given below.

- It proposes a novel feature generation method for a Facebook dataset by using Relief and NCA to generate weights passed through an activation function to arrive at the final feature vector. Though NCA and Relief are primarily for feature selection, they are employed for feature generation in this work.
- Many methods in machine learning assign weights randomly and then use the weight updating method to optimize the weights for high performances. Unfortunately, this process

increases the training time of the model. However, the application of NCA and Relief generates the weights without any need for weight updating and optimization techniques. Thus, making the implementation time shorter.

- The proposed model showcases the efficiency of regression models on the FB dataset and the likes with high prediction accuracy.
- This proposed method of regression is suitable for social media monitoring

2. Materials and Method

This work uses the Facebook dataset. The dataset comprises posts published between January 2014 and December 2014. It consists of 500 rows and 19 columns. The columns are used as features, with nine columns out of the 19 having continuous values. These nine attributes are used as the output target in each experiment, one after the other with the proposed method.

2.1 The Proposed Method

The proposed method consists of data normalization, weights generation using NCA or ReliefF, feature calculation using activation function plus weights together, and regression phases. The general graphical outline of the proposed method is shown in Fig 2 below.

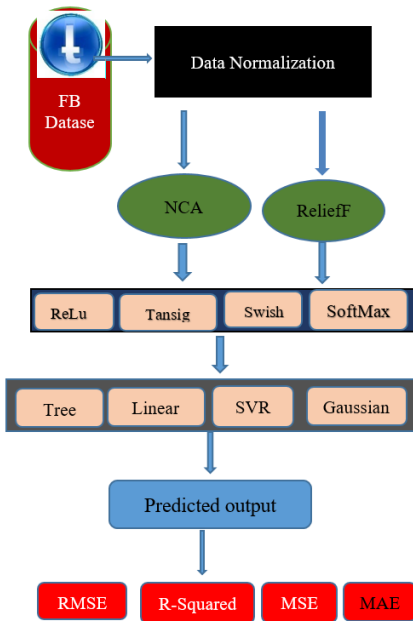


Figure 2. Graphical representation of the proposed method

In the first step, the min-max normalization method transforms the dataset into an appropriate form for feature generation. ReliefF and NCA first generate weights separately. Next, the generated weights are passed through an activation function to generate the desired features. Finally, these features get fed to the regression models for prediction.

The experiment is performed separately for the two methods (ReliefF & NCA). The experiment is first

conducted using ReliefF as the weight generation method, after which it is carried out using NCA as the weight generation method. The regression models used are Linear Regression, Tree, SVR, and Gaussian process regression. To test the model, all the nine continuous labels in the data were predicted one after the other. Thus, 9 case instances were created from predicting these attributes.

As a result, eight different results were obtained for each case scenario, i.e., two sets of results for each regression model based on the two weight generation methods of ReliefF and NCA. Figure 3 and 4 show schematic explanation of ReliefF and NCA.

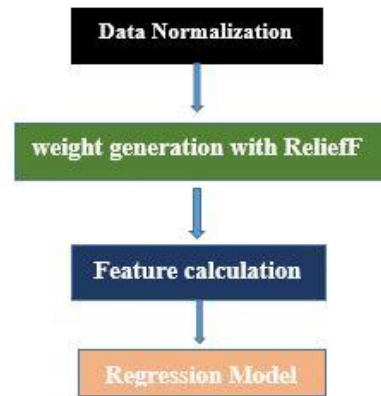


Figure 3. Schematic explanation of the model based on ReliefF weights generation method

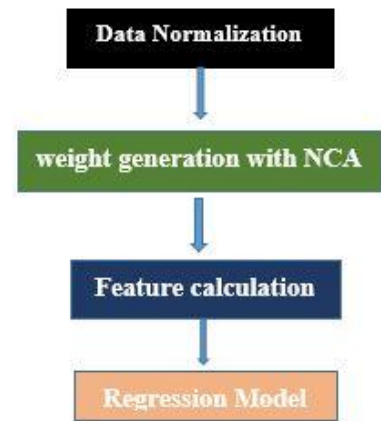


Figure 4. Schematic explanation of the model based on NCA weights generation method

2.1.1. Normalization and data separation

The dataset has 19 columns (attributes), out of which nine columns are continuous data. In the first step, we normalized these values by using the min-max normalization method. The mathematical description of the min-max norm is given in Equation 1 below.

$$D^N = \frac{D - \min(x)}{\text{Max}(D) - \text{Min}(D)} \quad (1)$$

where D is raw data, D^N Normalized data, $\min(\cdot)$ and $\max(\cdot)$ are the minima and maximum functions.

2.1.2. Weight Generation

Before calculating the features, we first applied a weight generation technique. The generated weights will be passed through an activation function before arriving at the final features. Next, ReliefF and neighborhood component analysis (NCA) methods are used to generate the weights. Finally, the weights from each technique are used to obtain the features used with the regression models to predict the class targets.

ReliefF handles samples from the data set and performs the attribute selection process by creating a model based on its proximity to the other examples in its class and its distance from different labels. It has been widely used for feature selection, and it generates weights of the features for feature ranking. The ReliefF method generates both negative and non-negative weights (Sun and Li 2006). The mathematical expression of the relief function is given as:

$$w^R = reliefF(X, T) \quad (2)$$

where w^R are weights of the ReliefF, $reliefF(\cdot)$ is the ReliefF-based weights generation function, $relief(\cdot)$

NCA: It is a non-parametric feature weighting method, and it generates non-negative weights for each feature. It generates feature weights using target values in the same way as Relief. This method uses distance calculation as in the 1-NN algorithm. Determination of the reference point according to a probability distribution is an effective method in the NCA [23].

1-NN: k-Nearest Neighbors (KNN) algorithm is a parametric classifier. We can use several parameters, for instance, distance metric and k value. If the k value selected is 1, this classifier is called 1-NN. The 1-NN is used in feature selection methods such as neighborhood component analysis (NCA) and Relief to calculate feature weights [24, 25].

The mathematical definition of NCA feature generation is given in Equation (3) below.

$$w^{NCA} = NCA(X, T) \quad (3)$$

where w^{NCA} Denotes the weights generated and $NCA(\cdot)$ is NCA-based weights generation function.

2.1.3. Feature calculation

Feature generation is an essential phase in this research. This step uses the generated weights from the prior phase with four activation functions to calculate the feature values. These weights are passed as inputs to the activation functions, which gives the final feature vector. The activation functions used are tangent sigmoid, swish, ReLu, and Softmax. The mathematical expressions for these functions are below (Balestriero and Baraniuk 2018; Lin and Wang 2008; Nwankpa et al. 2018).

Tangent Sigmoid:

$$\sigma(y) = \tanh(y) = \frac{e^y - e^{-y}}{e^y + e^{-y}} \quad (4)$$

Swish:

$$\sigma(y) = Swish(y) = \frac{y}{1 + e^{-y}} \quad (5)$$

ReLu:

$$\sigma(y) = ReLu(y) = \begin{cases} 0, & y < 0 \\ y, & y \geq 0 \end{cases} \quad (6)$$

Softmax:

$$\sigma(y) = \frac{e^{y_k}}{\sum_{k=1}^J e^{y_k}}, k = \{1, 2, \dots, J\} \quad (7)$$

where $\sigma(\cdot)$ Defines activation function and y is an input parameter to it. The listed activation functions are extensively used in machine and deep learning models. Hence, we apply them here for feature calculation.

2.1.4. Regression

This is the final step of the proposed method. Regression is a statistical method for finding correlations between a dependent variable and one or more independent variables. In this step, linear, tree, SVR, and Gaussian regression models are employed to obtain the results. Regression models are primarily used on datasets whose values are continuous. A brief description of these models is given below.

Linear regression: Linear regression is one of the most commonly used regressors in the literature, and it is the simplest regressor. A regression method models the relationship between two or more variables using a linear equation [26].

$$Y(X) = W_0 + W_1X_1 \dots + W_nX_n \quad (8)$$

Decision tree regression: Tree is one of the most commonly used algorithms for supervised learning. It has various versions, and it has been used in both classification and regression tasks. For this work, we used the medium tree model in MATLAB with a minimum leaf size of 12 [27].

Support vector regression (SVR): A class of an SVM model used for a regression task. It maps the original input data points to a higher dimensional feature of space where an optimal separating hyperplane is built. The algorithm can use a different set of mathematical functions called kernel functions. We used the quadratic kernel function in this work, which gave the highest performance [28]. The mathematical expression of the SVR model is given below.

$$f(x) = W^T \varphi(x) + b \quad (9)$$

where $w \in R^D$ is a weight vector, T stands for the transpose operator. b is a bias, φ is a non-linear transfer function mapping the input vectors into a high dimensional feature space.

Gaussian process regression (GPR): The Gaussian process is a phenomenon in statistics used in the regression. It uses the distribution of the data, and there are many Gaussians-based regressors in MATLAB. We

have chosen the Gaussian-based regressor with a preset of 5/2 GPR and constant basis function [29].

The pseudo-code of the proposed method below helps in a better understanding of the model.

Algorithm 2. Pseudo-code of the proposed method.

<p>Input: Input values (X) with size of 500 x 18 and target values (T) with size of 500 x 1</p> <p>Output: Results (r)</p>
<p>1: Normalize X and T values using Eq. (1). 2: Select ReliefF or NCA to generate weights. 3: Generate weights using Eq. (2) or Eq. (3). 4: Calculate features using the weight and the selected activation functions 5: Obtain results using linear, SVR, tree, and Gaussian regressions. 6: Evaluate the performance of the model using the selected evaluation metrics.</p>

3. Performance Analysis and Results

The tests are implemented on a personal computer with 16 gigabytes (GB) RAM, intel i7-7700 CPU with 3.60 GHz on Windows 10.1 operating system. The simulations are programmed by MATLAB2018a and the regression learner toolbox of the MATLAB2018a. The linear regression, tree, SVR, and GPR methods are used as regressors.

To evaluate the performance of the proposed method for regression, 9 case scenarios were established. In each case scenario, a continuous output column is selected and set as the output target. The predictions for each case scenario are made using the four regression models, first with features generated from the ReliefF model and then with features generated from the NCA method. Therefore, eight results are generated per case scenario. In addition, Root Mean Square Error (RMSE), R-square, mean square error (MSE), mean absolute error (MAE), and training times are chosen for numerical performance evaluation [30-32]. The mathematical definitions of these metrics are given in Eq. (10) to (13).

$$RMSE = \sqrt{MSE} \quad (10)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (r_i - p_i)^2 \quad (11)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |r_i - p_i| \quad (12)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (r_i - p_i)^2}{\sum_{i=1}^N (r_i - \bar{r})^2} \quad (13)$$

where R^2 defines R-square, N is the length of the output, r represents actual outputs, \bar{r} is the average value of the

actual outputs, and p is the predicted output. The defined cases and their results are given below.

3.1. Case 1

In this case, the page total likes column (column 1) is used as a target, and other columns are used as inputs. The best results obtained with each feature generation method were recorded and listed in Table 1 below.

Table 1. Results obtained for Case 1.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	Tree	Softmax	0.024	0.99			1.466
NCA	Tree	Tansig	0.024	0.99			1.5291

The graphical summarization of best results for Case 1 items is shown in Fig 5.

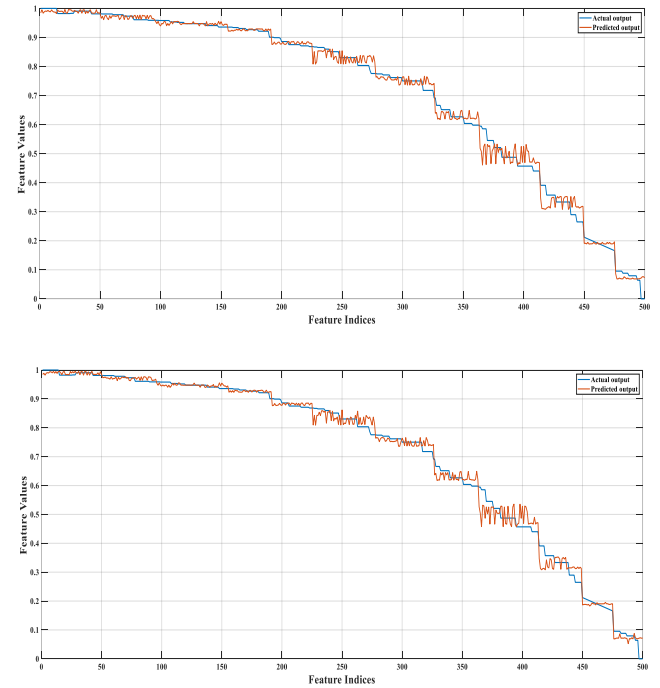


Figure 5. The actual outputs of items 1 & 2 case 1.

3.2 Case 2

The experiment, in this case, is performed by setting the lifetime post total reach (8th column) of the dataset as the target. The best results for this case are listed in Table 2.

Table 2. Results obtained for Case 2.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	Linear Regression	Softmax	0.0359	0.92	0.0012	0.0155	3.2578
NCA	Linear Regression	ReLU	0.0384	0.91	0.0014	0.016	3.5444

The graphical summarization of best results for Case 2 items is shown in Fig 6 below.

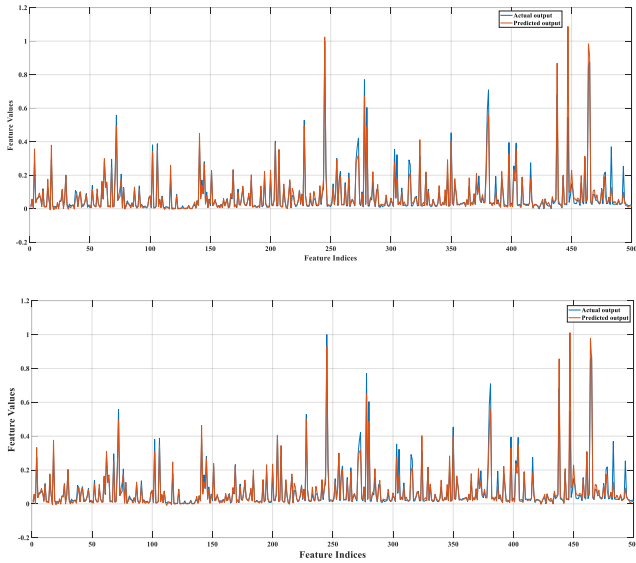


Figure 6. The actual outputs of items 1 & 2 of Case 2

3.3 Case 3

In this case, the Lifetime post total impressions column (9th column) is used as output. The results are listed in Table 3 below.

Table 3. Results obtained for Case 3.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	Linear Regression	ReLu	0.015312	0.95	0.00023	0.00626	3.1613
NCA	Linear Regression	ReLu	0.015841	0.95	0.00025	0.00642	3.1679

The output plots of Relief and NCA based weight generators for case 3 are shown in Fig 7 below.

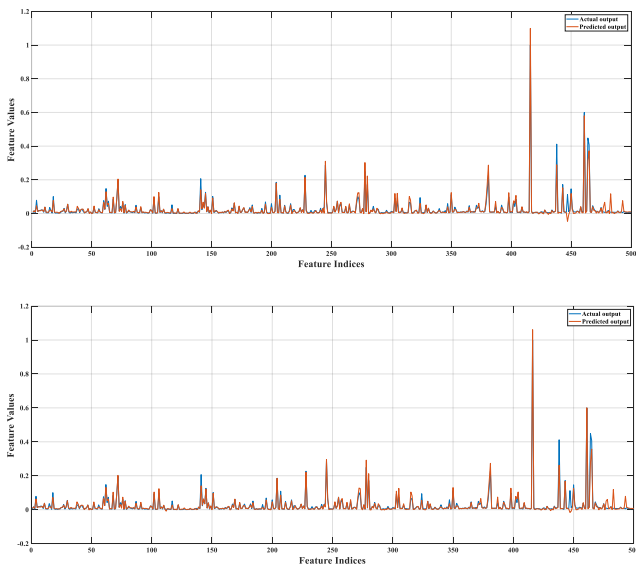


Figure 7. The actual outputs of items 1 & 2 of Case 3

3.4 Case 4

In Case 4, the lifetime engaged users are set as the target output. The obtained results from predicting this column are listed in Table 4.

Table 4. Results obtained for Case 4.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	Linear Regression	Swish	0.003529	1	0.0000124	0.00177	3.1652
NCA	Linear Regression	ReLu	0.004412	1	0.0000196	0.00184	3.1757

The output plots of Relief and NCA based weight generators for case 4 are shown in Fig 8 below.

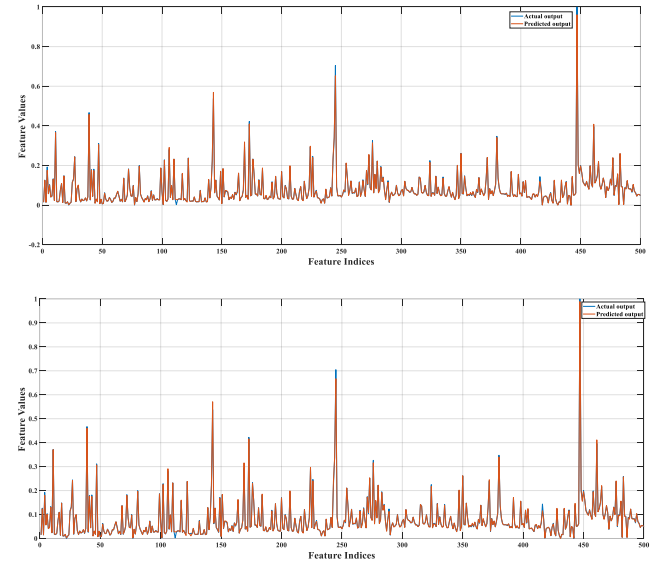


Figure 8. The actual outputs of items 1 & 2 of Case 4.

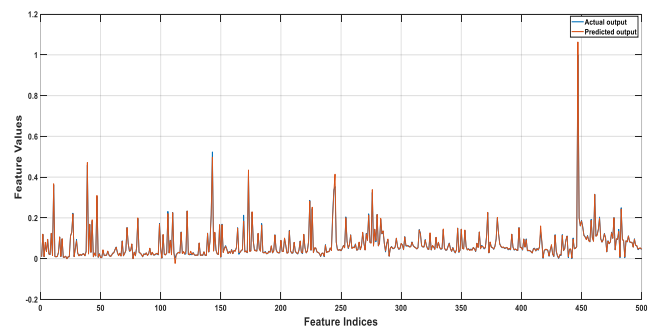
3.5 Case 5

In this experiment, the Life Post Consumers (11th column) column is used as output. The best results are listed in Table 5 below.

Table 5. Results obtained for Case 5.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	Linear Regression	ReLu	0.00413	1	0.0000170	0.00182	3.1717
NCA	Linear Regression	ReLu	0.00422	1	0.0000178	0.00184	3.2341

The output plots of Relief and NCA-based weight generators for case 5 are shown in Fig 9 below.



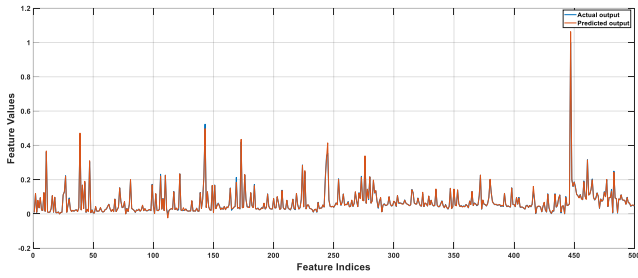


Figure 9. The actual outputs of items 1 & 2 of Case 5

3.5 Case 6

In this case, the column “Lifetime Post Impressions by people who have liked your Page” is used as a target. The results of this case are recorded in table 6 below.

Table 6. Results obtained for Case 6.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	SVR	ReLU	0.0065496	0.99	0.000042898	0.0023	2.4875
NCA	SVR	ReLU	0.0010284	0.96	0.00010576	0.0027	2.724

The output plots of Relief and NCA based weight generators for case 6 are shown in Fig 10 below.

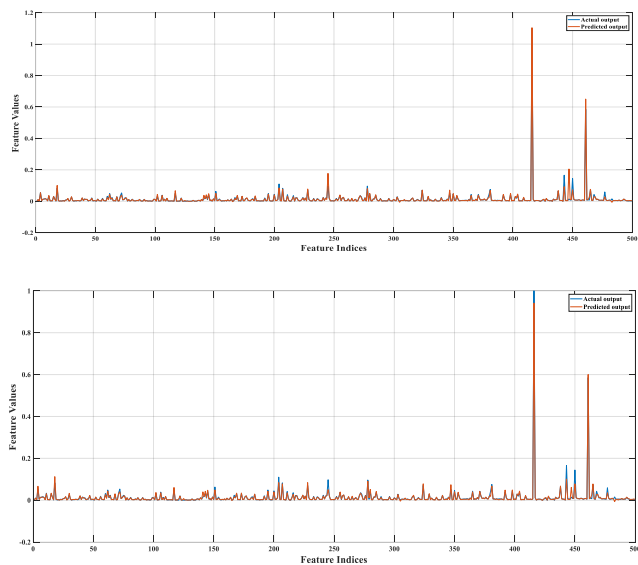


Figure 10. The actual outputs of items 1 and 2 of the Case 6

3.7 Case 7

In this case, the “Lifetime Post reach by people who like your Page” column (14th column) is used as output. The computed results of this case are listed in Table 7 below.

Table 7. Results obtained for Case 7.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	Tree	Tansig	0.054352	0.87	0.0029541	0.021344	2.8501
NCA	Linear Regression	Tansig	0.050182	0.89	0.0025182	0.026137	3.6412

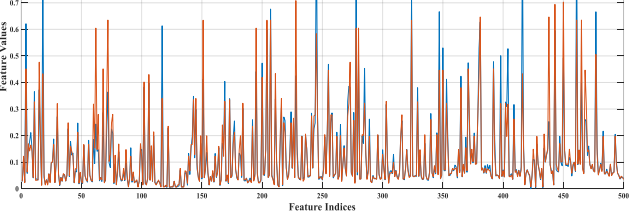
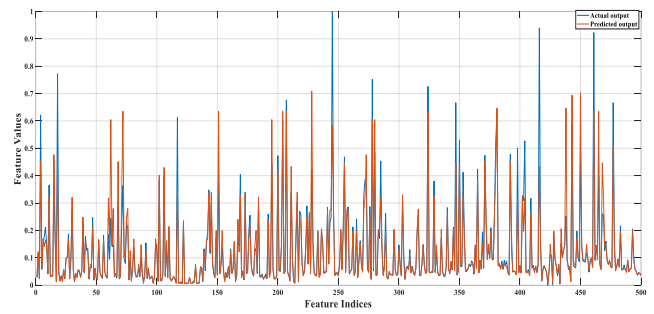


Figure 11. The actual outputs of items 1 & 2 of Case 7

3.8 Case 8

The “likes” column is utilized as output (17th column), and the other 18 columns are used as inputs. The results and plots for this case are given below.

Table 8. Results obtained for Case 8.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	Linear Regression	Tansig	0.00367	1	0.000135	0.001507	3.7256
NCA	Linear Regression	ReLU	0.00384	1	0.000015	0.001539	4.0079

The output plots of Relief and NCA based weight generators for case 8 are shown in Fig 12 below.

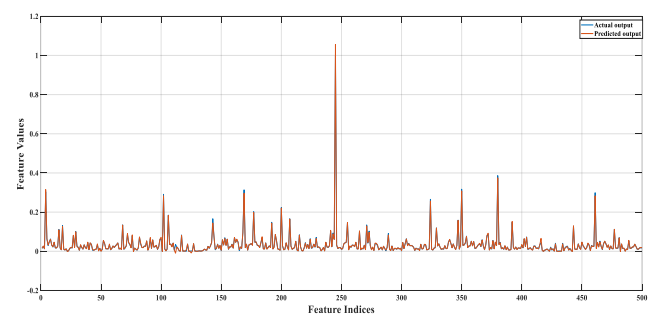
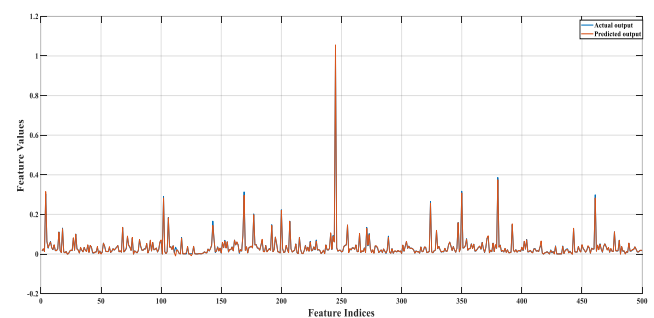


Figure 12. The actual outputs of items 1&2 of Case 8.

3.9 Case 9

Case 9 predicts “Total Interactions” using a regression model. In this case, the total interaction column (19th column) is used as output, and others are used as input. The best results are listed in Table 9.

Table 9. Results obtained for Case 9.

Feature Generation Model	Regression Model	Activation Function	Evaluation Metrics				
			RMSE	R-Squared	MSE	MAE	Training Time
Relief	Linear Regression	SoftMax	0.0028073	1	0.0000079	0.001468	4.1315
NCA	Linear Regression	ReLu	0.0040001	1	0.000016	0.001661	4.6203

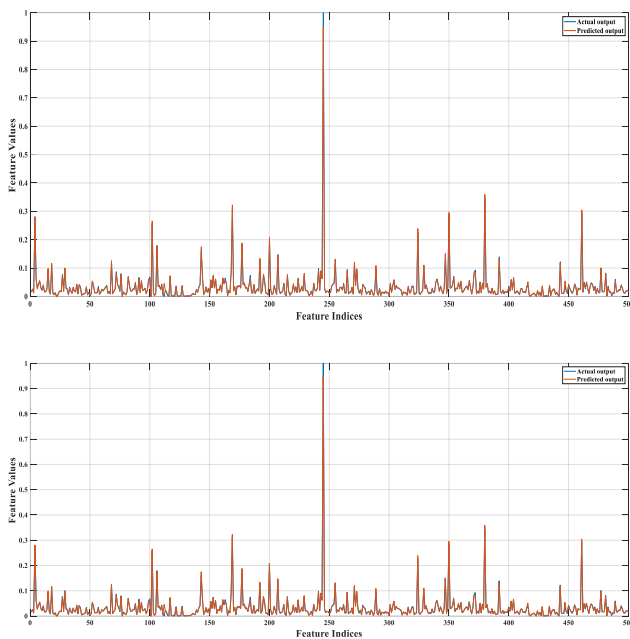


Figure 13. The actual outputs of items 1 & 2 of Case 9

4. Discussion

The results depict the performances of the different regressors, activation functions, and weight generators on a Facebook dataset. Evaluation metrics such as RMSE, R-Squared, MSE, and MAE were used to evaluate the performances of the proposed methods. For each case scenario, we first take the weights generated via the relief weight generation approach and use these weights with the various activation functions to calculate our features, then pass the features to the regression model. The process is also repeated for the NCA method. Finally, the best experimental results from the two operations are recorded alongside the regression model and activation function that yielded those results. The experiments clearly showed the effectiveness of our model as indicated by the metrics.

A comprehensive summary of the result is given below.

- The highest results are obtained for Case 4,5,8 and 9 because the R-squared value for these cases equals 1.
- The best activation function for the regression task on the Facebook dataset is ReLu. Out of 18

best-selected results, ten were obtained using ReLu as the activation function.

- This task’s best performing regression model is the linear regression model. 15 out of the 18 best results were obtained using linear regression.
- NCA-based features performed better compared to relief-based features in terms of prediction accuracy. ReliefF, however, has a shorter execution time.
- The worst R-squared value from the entire experiment is 0.87, which is satisfactory for a regression task.

The figure below shows the performance of the different activation functions and the regression models. It can be seen that the Gaussian process does not yield any top result and thus does not reflect on the chart.

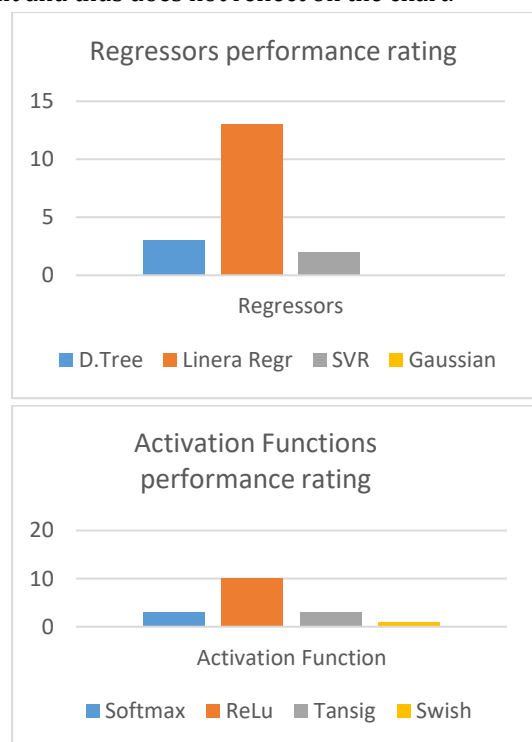


Figure 14. Performance analysis of regressors and activation functions

The method proposed here is simple with a simple mathematical background. It is, therefore, straightforward to implement with a short execution time. The weight generation techniques of NCA and Relief do not require weight updates or weight optimization, which contributes to the model’s short execution time. The proposed feature calculation method is novel for Facebook data. The use of four essential activation functions alongside the NCA and ReliefF weight generation method for feature calculation further improves the cognitive ability of the model. The experimental result demonstrates that the proposed method achieved high performance for all cases. Hence, we can conclude that our approach can extract universal informative features.

The limitation of the proposed method is in the amount of dataset used; we used a small database. This can be improved by extending the application to a more extensive database.

5. Conclusions

In this paper, a novel regression model is proposed to predict the Facebook data. The developed method consists of normalization, weight generation, feature calculation, and regression steps. This technique is a cognitive method as we have used NCA and relief weight generators. Four widely used activation functions are employed for feature extraction. The extracted features are forwarded to tree, linear, SVR, Gaussian process regressors. To prove the success of the proposed method, we used it to predict nine different attributes of the Facebook dataset. As a result, the model achieved an R-Squared value of 1 in predicting four different attributes (case 4,5,8 and 9) and reached an average R-Squared value of 0.9678 for all the predictions. These results demonstrate that the proposed method is efficient.

In the future, the proposed method can be applied for regression tasks on other datasets. Big datasets can be used with our developed regression model. The application can be extended to deep learning methods to predict social datasets.

6. Acknowledgment

The authors reported no potential conflict of interest.

References

- [1] Sutcliffe, A. G., Binder, J. F., and Dunbar, R. I., "Activity in social media and intimacy in social relationships," *Computers in Human Behavior*, vol. 85, pp. 227-235, 2018.
- [2] Zeppelzauer, M. and Schopfhauser, D., "Multimodal classification of events in social media," *Image and Vision Computing*, vol. 53, pp. 45-56, 2016.
- [3] Petkos, G., Papadopoulos, S., and Kompatsiaris, Y., "Social event detection using multimodal clustering and integrating supervisory signals," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 2012, p. 23.
- [4] Petkos, G., Papadopoulos, S., Mezaris, V., and Kompatsiaris, Y., "Social Event Detection at MediaEval 2014: Challenges, Datasets, and Evaluation," in *MediaEval*, 2014.
- [5] Yadav, M., Joshi, Y., and Rahman, Z., "Mobile social media: The new hybrid element of digital marketing communications," *Procedia-social and behavioral Sciences*, vol. 189, pp. 335-343, 2015.
- [6] Atzori, L., Iera, A., Morabito, G., and Nitti, M., "The social internet of things (siot)-when social networks meet the internet of things: Concept, architecture and network characterization," *Computer networks*, vol. 56, pp. 3594-3608, 2012.
- [7] Batrinca, B. and Treleaven, P. C., "Social media analytics: a survey of techniques, tools and platforms," *Ai & Society*, vol. 30, pp. 89-116, 2015.
- [8] Marturana, F. and Tacconi, S., "A Machine Learning-based Triage methodology for automated categorization of digital media," *Digital Investigation*, vol. 10, pp. 193-204, 2013.
- [9] Dey, N., Borah, S., Babo, R., and Ashour, A. S., *Social Network Analytics: Computational Research Methods and Techniques*: Academic Press, 2018.
- [10] Raynes-Goldie, K., "Aliases, creeping, and wall cleaning: Understanding privacy in the age of Facebook," *First Monday*, vol. 15, 2010.
- [11] Singh, M., Bansal, D., and Sofat, S., "Behavioral analysis and classification of spammers distributing pornographic content in social media," *Social Network Analysis and Mining*, vol. 6, p. 41, 2016.
- [12] Injadat, M., Salo, F., and Nassif, A. B., "Data mining techniques in social media: A survey," *Neurocomputing*, vol. 214, pp. 654-670, 2016.
- [13] Sapountzi, A. and Psannis, K. E., "Social networking data analysis tools & challenges," *Future Generation Computer Systems*, vol. 86, pp. 893-913, 2018.
- [14] Panigrahi, R. and Borah, S., "Classification and Analysis of Facebook Metrics Dataset Using Supervised Classifiers," *Social Network Analytics: Computational Research Methods and Techniques*, p. 1, 2018.
- [15] Cui, Y., Meng, C., He, Q., and Gao, J., "Forecasting current and next trip purpose with social media data and Google Places," *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 159-174, 2018.
- [16] Zhang, Z., He, Q., Gao, J., and Ni, M., "A deep learning approach for detecting traffic accidents from social media data," *Transportation research part C: emerging technologies*, vol. 86, pp. 580-596, 2018.
- [17] Ertugrul, Ö. F., "Forecasting electricity load by a novel recurrent extreme learning machines approach," *International Journal of Electrical Power & Energy Systems*, vol. 78, pp. 429-435, 2016.
- [18] Fernández-Delgado, M., Sirsat, M., Cernadas, Alawadi, E., S., Barro, S., and Febrero-Bande, M., "An extensive experimental survey of regression methods," *Neural Networks*, 2018.
- [19] Vanli, N. D., Sayin, M. O., Mohaghegh, M., Ozkan, H., and Kozat, S. S., "Nonlinear regression via incremental decision trees," *Pattern Recognition*, vol. 86, pp. 1-13, 2019.
- [20] Van Erp, S., Oberski, D. L., and Mulder, J., "Shrinkage priors for Bayesian penalized regression," *Journal of Mathematical Psychology*, vol. 89, pp. 31-50, 2019.

- [21] Ertuğrul, Ö. F. and Tağluk, M. E., "A novel version of k nearest neighbor: Dependent nearest neighbor," *Applied Soft Computing*, vol. 55, pp. 480-490, 2017.
- [22] Prashanth, R., Roy, S. D., Mandal, P. K., and Ghosh, S., "Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging," *Expert Systems with Applications*, vol. 41, pp. 3333-3342, 2014.
- [23] Yang, W., Wang, K., and Zuo, W., "Fast neighborhood component analysis," *Neurocomputing*, vol. 83, pp. 31-37, 2012.
- [24] Oliva, J. T. and Rosa, J. L. G., "Classification for EEG Report Generation and Epilepsy Detection," *Neurocomputing*, 2019.
- [25] Alpaydin, E., *Introduction to machine learning*: MIT press, 2014.
- [26] Seber, G. A. and Lee, A. J., *Linear regression analysis* vol. 329: John Wiley & Sons, 2012.
- [27] Bui, D. T., Tuan, T. A., Klempe, H., Pradhan, B., and Revhaug, I., "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree," *Landslides*, vol. 13, pp. 361-378, 2016.
- [28] Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V., "Support vector regression machines," in *Advances in neural information processing systems*, 1997, pp. 155-161.
- [29] Hultquist, C., Chen, G., and Zhao, K., "A comparison of Gaussian process regression, random forests and support vector regression for burn severity assessment in diseased forests," *Remote sensing letters*, vol. 5, pp. 723-732, 2014.
- [30] Balestriero, R. and Baraniuk, R. G., "From Hard to Soft: Understanding Deep Network Nonlinearities via Vector Quantization and Statistical Inference," *arXiv preprint arXiv:1810.09274*, 2018.
- [31] Sharma, K., Garg, R., Nagpal, C., and Garg, R., "Selection of optimal software reliability growth models using a distance based approach," *IEEE Transactions on Reliability*, vol. 59, pp. 266-276, 2010.
- [32] Kanmani, S., Uthariaraj, V. R., Sankaranarayanan, V., and Thambidurai, P., "Object oriented software quality prediction using general regression neural networks," *ACM SIGSOFT Software Engineering Notes*, vol. 29, pp. 1-6, 2004.