Journal of Advanced Sciences and Nanotechnology

Journal homepage: https://joasnt.com

The Effect of Feature Selection Methods in Detecting Malicious Accounts in Social Media

Amna Kadhim Ali [1, *], Faten Salim Hanoon [2], Sabreen Fawzi Raheem [3]

[1] College of Veterinary Medicine, University of Basrah, Basrah, Iraq.

[2] Ministry of Education, General Directorate of Education, Basrah Gifted School, Basrah, Iraq.

[3] Basrah Technical Institute, Southern Technical University, Basrah, Iraq.

Article information

Article history:
Received: November, 11, 2022
Accepted: February, 23, 2023
Available online: April, 15, 2022

Keywords:
Feature Selection Methods,
Machine Learning,
Social Media,
Recursive Feature Elimination.

Corresponding Author:
Amna Kadhim Ali
amna.kadhim@uobasrah.edu.iq

Citation : Fawzi Raheem, S., Kadhim Ali , A., & Salim Hanoon , F. (2023). The Effect of Feature Selection Methods in Detecting Malicious Accounts in Social Media . Journal of Advanced Sciences and Nanotechnology, 2(1), 215–224.

Abstract

Social network is a significant aspect of our lives since it has become an essential part of global communication. Twitter is a microblogging network platform, with an average of more than half a billion tweets posted per day by millions of users. With such diversity and widespread use, Twitter is easily affected with malicious accounts by using fake profiles by hackers to carry out malicious activities, this network is suffering from identity theft via fake accounts. In this work, we focused on feature selection methods to improve the performance to detect these fake accounts, we propose decision tree (DT) - recursive feature elimination (DT-RFE) algorithm as a feature selection method to choose the best features in the personal profiles with data standardization to speed up processing, finally, the efficiency of computational detection was evaluated using a collection of supervised machine learning algorithms. The study showed a high degree of accuracy in classifying the accounts on all proposed algorithms using the feature selection method, with the highest degree of accuracy of 94% by using the random forest algorithm, and even showed the importance of feature selection methods in enhancing detection accuracy by reducing non-significant features that negatively affect the classification process.

DOI: https://doi.org/10.55945/joasnt.2023.2.1.215-224,
ISSN: 2791-0903/

1. Introduction

Nowadays, social networks are very much used, and people spend a lot of time on them [1], celebrities and big companies use networks to connect with their followers and send photos, messages, or even videos, news agencies also use these networks to broadcast news.

Along with the growing popularity and proliferation of online social networks, dangers and security threats have escalated, potentially jeopardizing user privacy and confidence by some illegal people who are trying to spread malicious activities with fake accounts [2].

Fake accounts can take many forms, and they are created in different ways, either by controlling some real accounts and exploiting them, or by cloning the personal files of real people and creating fake accounts in the same network or in another network, or by using fake information to create these types of accounts and use them for malicious purposes [3]. It is considered one of the disadvantages of social networking sites that cannot be controlled, but it is being addressed in various ways to reduce it.

Social media platforms such as Twitter, Facebook, and Instagram are among the most vulnerable to fake accounts [4]. Due to the high number of users as well as the simplicity of creating these accounts, it is a good environment that can be used for illegal purposes such as blackmail, defamation, or other malicious purposes.

Tweets from fake accounts that advertise spam websites or services are a waste of time and energy for everyone involved. Furthermore, harmful substances have become more widely available because of the ease with which fraudulent information may be sent to users under false identities. This false and fraudulent user detection has become a hot topic in social network studies recently [5].

Feature selection is the process of identifying a subset of the most relevant and significant features from a larger set of features to be used in a machine learning model [6]. The benefit of feature selection is to improve the performance of the model by reducing its complexity, reducing overfitting, and improving its interpretability.

The aim of this paper is to clarify the importance of using feature selection methods and the need to rely on them as a method for pre-processing data before using machine learning algorithms by demonstrating their role in improving detection accuracy and reducing execution time by reducing of unnecessary information in files. Because the primary goal for each algorithm is to achieve the highest performance.

In this work, we focused on the social networking platform Twitter, which has millions of active users and suffers from the spread of fake accounts. We propose the DT-RFE algorithm [7] as a feature selection method and standardization of the data as a preprocessing stage. And then we used a group of well-known machine learning algorithms to extract the accuracy and compared the result extracted using the proposed method with the result of the algorithms with all existing features.

2. Related Works

The detection and categorization of fake accounts using AI-assisted methods has gained significant attention from researchers all around the world. This section intends to review some of these previously conducted research that were used to identify fake accounts on online social networks and to give a summary of each study's methodology.

Using (supervised machine learning techniques), Suheel et al. [8] demonstrated a method for predicting fake Facebook profiles. Initially, the proposed model applied extensive noise removal and data normalization approaches, followed by the development of "Artificial Bee Colony (ABC)" and "Ant Colony Optimization (ACO)" to detect insignificant features in datasets and execute attribute reduction correspondingly. The proposed model was trained using an ensemble classifier .The evaluation of the theoretical model using ensemble classifiers revealed good performance using the “Weka tool” for detecting fake Facebook profiles.

Mohammadreza et al. [9] presented technique for identifying fake accounts in social networks. This method used the network graph to calculate the adjacency matrix. Furthermore, the adjacency matrix was used to determine the similarities. New characteristics were extracted by utilizing the elbow approach and principal component analysis. Using one-class algorithms, they trained a model

that accurately identified fake accounts. The accuracy and false negative rates for the Twitter dataset were 99.6% and 0%, respectively, according to experimental findings.

Adebola [10] established a mechanism in this study to successfully identify fake profiles in Online Social Networks (OSN). The author used natural language processing to delete or reduce the quantity of the dataset, hence enhancing the model's overall performance. Using principal component analysis, an appropriate selection of features was achieved. Six variables or features affecting the classifier were identified following extraction. As classifiers, we employed Support Vector Machine (SVM), Naive Bayes, and Improved Support Vector Machine (ISVM). ISVM introduces a penalty parameter for the normal SVM objective function to relax inequality restrictions between slack variables. This produced a superior result of 90% when compared to the SVM and Naive Bayes, which produced results of 77.4% and 77.3%, respectively.

Muhammad et al. [11] suggested a technique that helps in identifying the text in Amazon-based reviews as spam and non-spam by introducing a rule-based feature weighting method and utilizing a hybrid set of features (Opinion Spam, Opinion Spammer, and Item Spam), prioritize the spamicity features using revised feature weighting scheme. The experimental findings, which are presented in the forms of accuracy, precision, recall, and F-measure, demonstrate that the proposed system outperformed the comparing techniques.

Aliaksandr and Petr [12] developed a brand-new, cost-conscious method for removing spam from social networks. Two steps make up their suggested strategy. The cost of misclassification for the suggested model as well as the number of characteristics needed for spam filtering are both reduced in the first stage using multi-objective evolutionary feature selection. After that, the method applies cost-sensitive ensemble learning techniques acting as the base learners with regularized deep neural networks.

On two benchmark datasets, they showed that this approach is effective for social network spam filtering. They also showed that the proposed approach outperforms other popular social network spam filtering algorithms such as random forest, Nave Bayes, and support vector machine.

Tushaar et al.[13] clarified how to extract email content and behavior-based features, as well as which features are appropriate for detecting unsolicited bulk emails (UBEs) and how to choose the most discriminating feature set. Additionally, in order to effectively address the threat posed by UBEs, they supported a thorough comparative analysis using a number of effective machine learning methods.

Amna and Abdulhussein [14] proposed a machine learning strategy for detecting fake Twitter accounts based on a collection of publicly accessible Twitter features. These feature sets were developed based on information contained in the user profiles. To extract the features in the detection process, they tested two distinct feature selection strategies, and the stack ensemble method, which is based on four machine learning algorithms, had the most influence on improving the detection model. Initial study results for the authors indicate that by combining logistic regression as a meta classifier with random forest, SVM, and naive Bayes as base level classifiers, a stack ensemble technique can achieve success.

3. Experimental Procedure

In this study, we aid in the detection of fake Twitter accounts with the highest degree of accuracy by following a strategy to reduce the number of features in the user's profile by choosing the best groups that effect on whether the account is real or fake by using a feature selection method as well as reducing the processing time using data standardization and finally classification of accounts by using a set of well-performing algorithms.

The working plan that has been performed to detect the required features' set and classify them are described in steps, steps below present the working plan in details.

A. Dataset

The dataset obtained was manually compiled by Buket E. [15], which includes, respectively, 501 and 499 fake and actual accounts with 13 features. It is worth noting that these features are collecting via the Twitter API. In addition to the creation of three features by the researcher, namely Hashtags-average, URL average and Mentions-average, this brings the number of features to 16.

Some features of this data are textual, with true or false values, And due to the use of some algorithms that do not deal with textual data, so the first step after obtaining the data is to convert the values of those text fields to numeric, where the fields containing false were converted to 0 and true to 1 to deal with them in the subsequent steps, the description of these features is shown in table (1).

Table 1: The description of all features

Feature name	Description
Twitter_user	Indicates the presence or absence of a username in a profile
Followers_count	The total number of followers that each account has
Friends_count	Describe how many people each account is following.
Favourites_count	Describe how many times each user account's tweets have been liked over the course of the account's existence.
Profile_has_background_image	When true, it means that the user's submitted background image was used by the account.
Hashtags_average	Explain how many hashtags each user account has used in the last 20 tweets.
Mentions_average	In the last 20 tweets, the number of mentions the user has used.
URL_average	The number of URL links used in the last 20 tweets by the user
Description	Is the user-defined account description string?
Statuses -count	The overall sum of tweets sent by each person in his or her account.
Listed_count	Described how many public groups and lists each user account belongs to
Verified	When true, it means the user's account has been checked.
Contributors	When the "contributor mode" is allowed, tweets sent by one account may be coauthored by another account.
Default_profiles	When true, it means the user hasn't changed his user profile's theme or context.
Default_profile_image	If this is so, it means that the account's owner hasn't posted a profile image, so the default one is used instead.
Translator	If this is accurate, it indicates that the user is a member of Twitter's translator network.

B. Feature Selection

To extract the significant features, features ranking with recursive feature elimination was employed to select the best features. A feature selection technique called recursive feature elimination (RFE) removes the weakest feature (or features) from a model until the required number of features is reached [16].

The information is obtained from the derived significance of the machine learning model, and it removes the feature only once every step. By eliminating features one at a time until the ideal amount of features is left, it minimizes model complexity in this manner.

In our proposed method we used REF with cross validation and decision tree DT to eliminate irrelevant features based on validation scores. DT is one of the most commonly used supervised learning algorithms, General DT virtualization is made up of multiple nodes specifically, root and leaf nodes that represent many classes [17].

The feature selection process includes training and testing the decision tree algorithm on all the data features and extracting the accuracy to evaluate its performance, (k-fold cross validation with 5 folds is used to splitting data to training and testing sets, where 4 folds are used as training data and the fifth one is used as testing data) then determine feature importance to rank features appropriately, remove the least important feature, then retrain the model on the remaining features while using the prior evaluation measure (accuracy) to determine the effectiveness of the final model, And to determine whether the assessment measure falls below a predetermined level, which indicates that this

feature is crucial. Otherwise, it will be removed (in our proposed method the threshold is auto to choose feature importance from feature importance of the decision tree algorithm). These steps are repeated until all features are removed. The flowchart below summarizes the work of the algorithm.

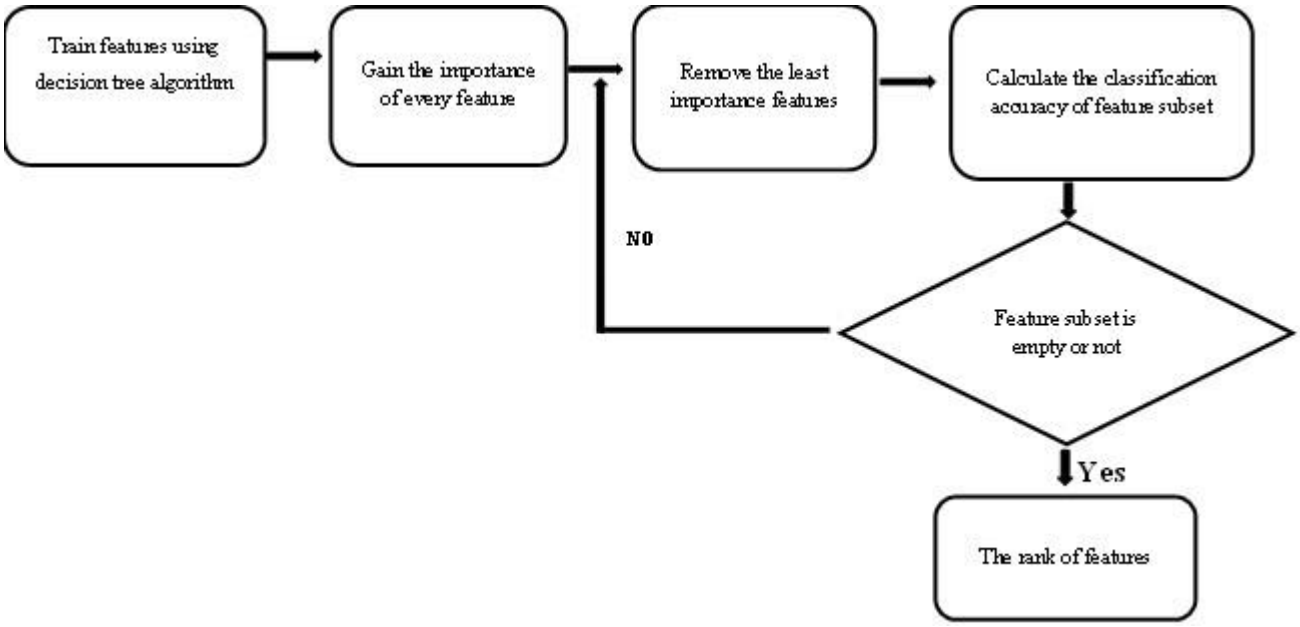


Figure1: The flowchart of DT-RFE algorithm

C. Data standardization

In order to accelerate the classification process by making the feature values in one level, data scaling is also performed at this stage. Standardization is one of the preprocessing procedures that scales each data input variable by subtracting the mean and dividing by the standard deviation, so reshaping the distribution to have a mean of zero and a standard deviation of one. This is the same as removing the mean or centering the data [18]. A value is standardized as follows:

$$\mathbf{y} = ((\mathbf{x} - \mathbf{mean}))/(\mathbf{standard\_deviation}) \quad (1)$$

Where the mean is calculated as:

$$\mathbf{mean} = \mathbf{sum}(\mathbf{x})/\mathbf{count}(\mathbf{x}) \quad (2)$$

And the standard-deviation is calculated as:

$$\mathbf{standard\ deviation} = \mathbf{sqrt}\ (\mathbf{sum}\ (\mathbf{x} - \mathbf{mean})^{\mathbf{2}})/\mathbf{count}(\mathbf{x})) \quad (3)$$

D. Classification

The process of classifying accounts into true and false was done by depending on supervised machine learning algorithms, five of the most widely used algorithms in the field of account detection, according to previous works that has been relied upon, namely: random forest, k-nearest neighbor, Naïve Bayes, support vector machine, and logistic regression.

- **Random Forest**

The Random Forest (RF) classifier is made up of a group of tree-shaped classifiers. It is a more complicated version of bagging [19] that uses randomization. RF divides each node by the best split among all variables, not by the best split among a random sample of predictors at that node. To make a new training data set, a replacement from the original data set is used. Then, a tree is grown by picking features at random. RF is also very fast, doesn't get too good at fitting, and can build as many trees as the user wants [20].

- **K-nearest neighbor**

The k-nearest neighbor method (KNN) is a strategy for categorizing objects in pattern recognition or classification that is based on the closest training samples in the problem space. KNN is an instance-based or lazy learning method in which the function is only estimated locally and full computation is postponed until classification [21]. The k-nearest neighbor method is one of the most basic machine learning algorithms: an item is categorized by a majority vote of its neighbors, with the object allocated to the class most prevalent among its k closest neighbors (k is a positive integer, typically small). If k is equal to one, the object is simply assigned to the class of its closest neighbor [22].

- **Support vector machine**

Support vector machine (SVM) is a popular machine learning method that has been used to solve pattern recognition problems. SVM is a kind of supervised machine learning. SVM training approach builds a model that predicts the category of a new sample given a series of training examples, each labeled as belonging to one of the numerous categories. SVM is better at generalizing problems, which is the purpose of statistical learning [23]. The primary idea underlying SVM is to build an ideal hyper plane that can be utilized for classification of linearly separable patterns. The ideal hyper plane is a hyper plane chosen from the set of hyper planes for pattern classification that maximizes the hyper plane's margin, or the distance from the hyper plane to the nearest point of each pattern. SVM's major goal is to maximize the margin so that it can properly categorize the provided patterns, i.e. the higher the margin size, the more accurately it classifies the patterns [24].

- **Naïve Bayes**

Naïve Bayes is a popular machine learning tool for classification, due to its simplicity, high computational efficiency. It is s a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable [25].

- **Logistic regression**

The logistic regression technique examines the relationship between one or more predictor factors and a response variable. Although regression provides superior results for numerical data values, it also allows for the prediction of discrete variables using a combination of continuous and discrete predictors. The functionality of discriminant function analysis and multiple regressions is the same, but there are no distributional assumptions on the resulting predictors. The predictors are linearly connected rather than normally distributed, and the variance is same in all groups [26].

The detection process included two experiments, first the algorithms were applied to all existing features without any processing and extract the accuracy, and the second experiment, unnecessary features were eliminated using the proposed feature selection method and scaling them, and then the algorithms were applied to the new data set.

Using stratified sampling [27] to ensure an equal division and maintaining the same proportion of classes in the train and test sets that are found in the entire original dataset, the data for both experiences were divided into training and test groups by selecting 75% of them as training data and 25% of them as testing data. Default parameters for all the algorithms were used.

E. Evaluation

A confusion matrix will be used as the main source of evaluation to assess the false detection models in the suggested system [28]. The accuracy and f1 score of the extracted results without preprocessing are discussed, and these results are compared with those obtained by using the feature selection method. Confusion matrix is appropriate to use when each class contains an equal number of observations, classification accuracy is defined as the number of accurate predictions as a percentage of all predictions made. The F-measure represents the harmonic mean of recall and precision. The ratio of positive predicted objects to positive actual objects is called precision. The ratio of positive outcomes to those that the system predicts would be favorable is known as recall.

4. Results and Discussion

In the present work, two scenarios have been studied. In the first scenario, all the features in the dataset have been employed in the training of the following machine learning algorithms random forest, KNN, SVM, Naïve Bayes and logistic regression. The results of this experiment that are representing with measurement of accuracy and f-score and execution time is shown in the table (2). The random forest proved its strong performance with binary-class classification compared with others algorithms. In terms of training time, it consumed a longer time, and the reason was due to the complex structure of the algorithm that made its training time is longer whereas Naïve Bayes consumed time less than the others algorithms.

Table 2: Applying all features

Algorithms	TP	TN	FP	FN	Accuracy	F_score	Time (seconds)
Random forest	121	113	12	4	93.6%	93.8%	1.678
Knn	103	94	31	22	78.8%	79.5%	1.048
Svm	119	16	109	6	54%	67.4%	1.064
Naïve bayes	118	44	81	7	64.8%	72.8%	1.023
Logistic regression	108	90	35	17	79.2%	80.6%	1.040

While the second scenario consists of two stages, the first is application of (DT-RFE) algorithm with data standardization, and as a result, the number of features is reduced to 11, which are statuses_count ,friends_count, hashtags_average , mentions_average , urls_average , description , followers_count ,favourites_count listed_count , default_profile_image , translator . Then the algorithms were applied to these extracted features and the result is shown in a table (3).

Table 3: The results after applying feature selection

Algorithms	TP	TN	FP	FN	Accuracy	F_score	Time (seconds)
Random forest	120	115	10	5	94%	94.8%	1.182
Knn	104	117	8	21	88.4%	87.8%	1.028
Svm	96	124	1	29	88%	86.4%	1.029
Naïve bayes	97	124	1	28	88.4%	86.9%	1.015
Logistic regression	101	118	7	24	87.6%	86.7%	1.030

As shown in Table(3), we see that the DT-RFE feature selection algorithm proved its efficiency and effectiveness in classification when it was used with machine learning algorithms, where from the results we notice a clear and very noticeable improvement in the criteria values when compared with the previous scenario. This is due to the selected useful features that help classification algorithms to effectively distinguish between fake users and legitimate users after using the feature selection method and deleting features that are repetitive or do not affect the type of account and thus negatively affect the classification result as in the first scenario.

In addition to the improvement of the criteria values, there is an improvement in the training time, especially for the random forest algorithm so that when we compare with the first scenario, we find that it consumed less training time as shown in the figure (2), and the reason is due to the low number of features, which led to fewer operations Arithmetic and fewer parameters, while the rest of the algorithms there is no significant difference in execution time, but the difference is very clear in terms of accuracy. The reason for the difference in training time is due to the complexity of the

calculations of the random forest algorithm compared to the rest of the algorithms.

In order to clearly observe the comparison results of detailed information, the figure (3) shows the accuracy and F-score obtained by the model to better explain the difference between each criterion. We can observe Random forest after applying feature selection .It has achieved accuracy, F_score of 94%, 94.8% respectively and consumed training time only 1.182s. Additionally, we are able to note the rest of the algorithms that were used and employed in this work improved their results after reducing the features, as Knn , Svm, Naïve bayes and Logistic regression have achieved accuracy of 88.4%, 88%, 88.4% , 87.6% and they have attained F_score of 87.8%,86.4%,86.9% and 86.7% respectively. As for the training time, it was as follows Knn, Svm, Naïve bayes and Logistic regression have taken time to train 1.028s, 1.029s, 1.015s and 1.030s. Using the features selection method has aided to focus only important features, which have contributed to a better results of model proposed.

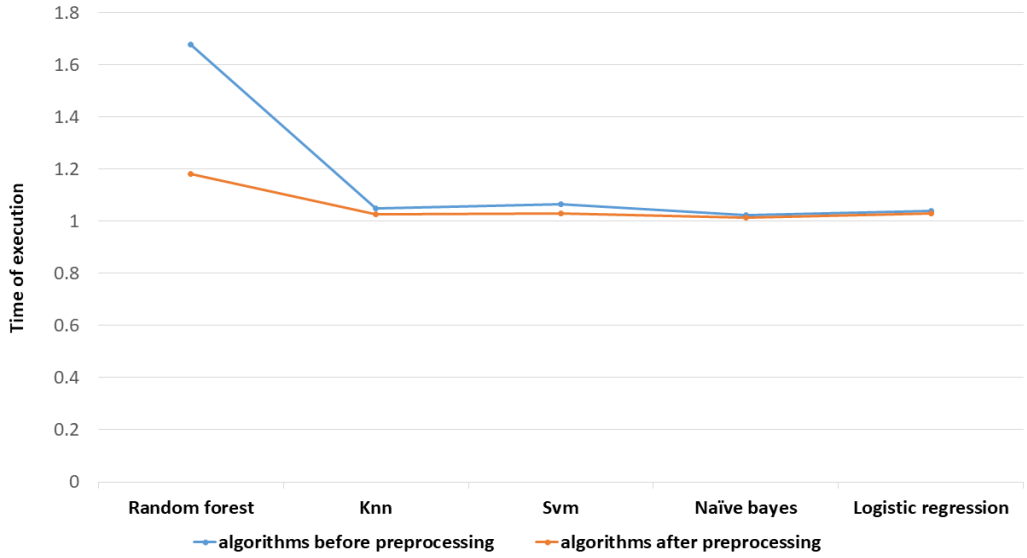
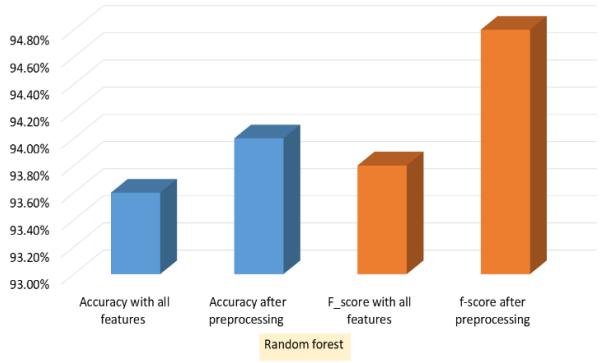
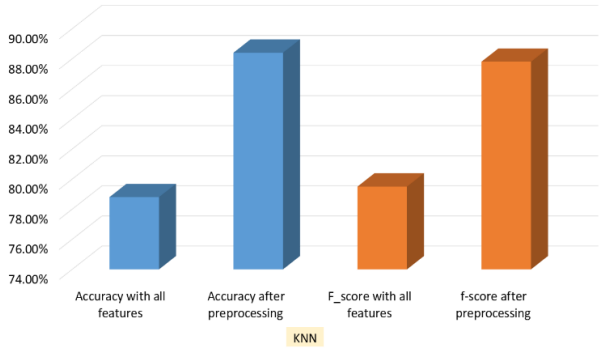


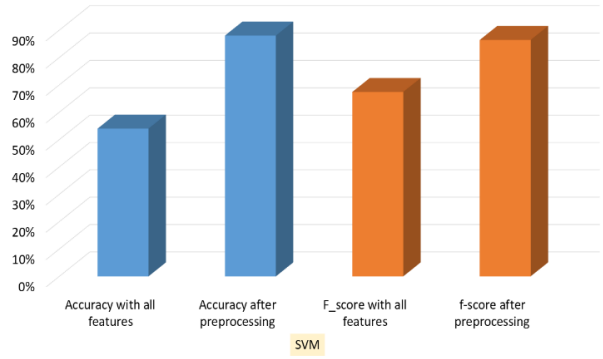
Figure 2: Execution time between algorithms after and before preprocessing



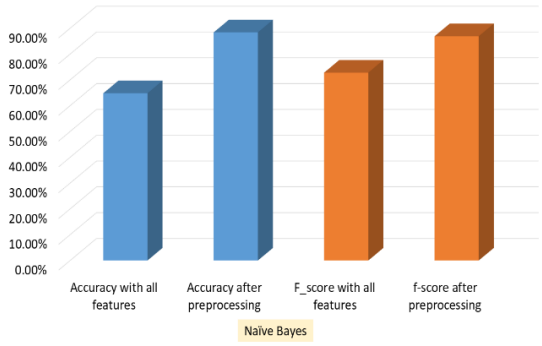
(a)



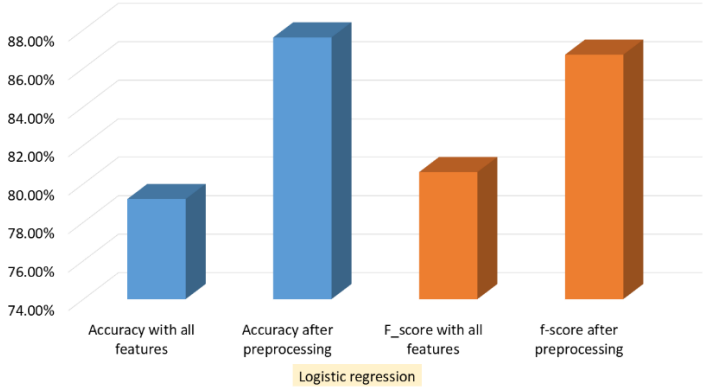
(b)



(c)



(d)



(e)

Figure 3: Visualization of the accuracy and f1-score for ((a) Random forest, (b) KNN, (c) SVM, (d) Naïve Bayes, (e) Logistic regression) between the first and second experiment

5. Conclusions

Fake accounts are one of the most important problems facing social media platforms because they may change concepts and harm the users of these platforms.

This study demonstrated the importance of feature selection methods as an initial stage for detecting fake accounts by proposing (DT-RFE) algorithm as a pre-processing of data to reduce the number of features in profiles, get rid of redundant and unimportant features, and use a set of machine learning algorithms for classification.

The proposed method showed high accuracy results for all algorithms used with using features after reduction compared to using these algorithms with all the features with the highest accuracy rate of the random forest algorithm 94%, while the knn algorithm gave an accuracy rate 88.4%, the Svm algorithm gave an accuracy rate 88%, the Naïve bayes algorithm gave an accuracy rate 88.4%, the Logistic regression gave an accuracy rate 87.6%. As all algorithms gave a clear difference in the results compared to the use of each database.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Milroy L, Llamas C (2013) Social Networks [E-book]. Handb Lang Var Chang (2nd ed, pp 407–427) John Wiley Sons, Inc https//doi org/101002/9781118335598 ch19
2. Khaled S, El-Tazi N, Mokhtar HMO (2018) Detecting fake accounts on social media. In: 2018 IEEE international conference on big data (big data). IEEE, pp 3672–3681
3. Gurajala S, White JS, Hudson B, et al (2016) Profile characteristics of fake Twitter accounts. Big Data Soc 3:2053951716674236
4. Kim J, Uddin ZA, Lee Y, et al (2021) A systematic review of the validity of screening depression through Facebook, Twitter, Instagram, and Snapchat. J Affect Disord 286:360–369
5. Shu K, Sliva A, Wang S, et al (2017) Fake news detection on social media: A data mining perspective. ACM SIGKDD Explor Newsl 19:22–36
6. Shardlow M (2016) An analysis of feature selection techniques. Univ Manchester 1:1–7
7. Tyagi A, Singh VP, Gore MM (2022) Improved Detection of Coronary Artery Disease Using DT-RFE Based Feature Selection and Ensemble Learning. In: Advanced Network Technologies and Intelligent Computing: First International Conference, ANTIC 2021, Varanasi, India, December 17–18, 2021, Proceedings. Springer, pp 425–440
8. Wani SY, Kirmani MM, Ansarulla SI (2016) Prediction of fake profiles on Facebook using supervised machine learning techniques-a theoretical model. Int J Comput Sci Inf Technol 7:1735–1738
9. Mohammadrezaei M, Mohammad ES, Rahmani AM (2019) Detection of fake accounts in social networks based on One Class Classification
10. Ojo AK (2019) Improved model for detecting fake profiles in online social network: A case study of twitter. J Adv Math Comput Sci 33:1–17

11. Asghar MZ, Ullah A, Ahmad S, Khan A (2020) Opinion spam detection framework using hybrid classification scheme. Soft Comput 24:3475–3498
12. Barushka A, Hajek P (2020) Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks. Neural Comput Appl 32:4239–4257
13. Gangavarapu T, Jaidhar CD, Chanduka B (2020) Applicability of machine learning in spam and phishing email filtering: review and approaches. Artif Intell Rev 53:5019–5081
14. Ali AK, Abdullah AM (2022) Fake accounts detection on social media using stack ensemble system. Int J Electr Comput Eng 12:
15. Erşahin B, Aktaş Ö, Kılınç D, Akyol C (2017) Twitter fake account detection. In: 2017 International Conference on Computer Science and Engineering (UBMK). IEEE, pp 388–392
16. Ding X, Yang F, Ma F (2022) An efficient model selection for linear discriminant function-based recursive feature elimination. J Biomed Inform 129:104070
17. Najm IA, Hamoud AK, Lloret J, Bosch I (2019) Machine learning prediction approach to enhance congestion control in 5G IoT environment. Electronics 8:607
18. Kumari B, Swarnkar T (2020) Importance of data standardization methods on stock indices prediction accuracy. In: Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2018, Volume 1. Springer, pp 309–318
19. Breiman L (2001) Using iterated bagging to debias regressions. Mach Learn 45:261–277
20. Iwendi C, Bashir AK, Peshkar A, et al (2020) COVID-19 patient health prediction using boosted random forest algorithm. Front public Heal 8:357
21. Raikwal JS, Saxena K (2012) Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. Int J Comput Appl 50:
22. Chomboon K, Chujai P, Teerarassamee P, et al (2015) An empirical study of distance metrics for k-nearest neighbor algorithm. In: Proceedings of the 3rd international conference on industrial application engineering
23. Pisner DA, Schnyer DM (2020) Support vector machine. In: Machine learning. Elsevier, pp 101–121
24. Pradhan A (2012) Support vector machine-a survey. Int J Emerg Technol Adv Eng 2:82–85
25. Vembandasamy K, Sasipriya R, Deepa E (2015) Heart diseases detection using Naive Bayes algorithm. Int J Innov Sci Eng Technol 2:441–444
26. Domínguez-Almendros S, Benítez-Parejo N, Gonzalez-Ramirez AR (2011) Logistic regression models. Allergol Immunopathol (Madr) 39:295–305
27. Ala'M A-Z, Alqatawna J, Paris H (2017) Spam profile detection in social networks based on public features. In: 2017 8th International Conference on information and Communication Systems (ICICS). IEEE, pp 130–135
28. Marom ND, Rokach L, Shmilovici A (2010) Using the confusion matrix for improving ensemble classifiers. In: 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel. IEEE, pp 555–559