Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.

Fake Account Detection using Machine Learning and Data Science

Rajashekar Nennuri[1], M Geetha Yadav[2], B. Shara[3], G. Anil Kumar[4], M. Shivani[5]
[1,2,3,4,5]Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad - 500043
rajasekharnennuri@gmail.com[1] geethayadav22@gmail.com[2], smileyshara00@gmail.com[3] , anilkumargarnepally938@gmail.com[4], maya.shivani@gmail.com[5]

ABSTRACT:

In today's world, Online Social Media is king in a number of forms the number of people who use the service is growing every day. The use of social media is skyrocketing. The primary benefit is that we can easily communicate with people via online social media and communicate with them in a more effective manner. This opened up a new avenue of a possible attack, such as a forged identity, false information and so on. According to a recent study, the number of accounts in the number of people who use social media is much higher than the number of people who use it. These fake accounts are difficult to detect for online social media providers. Since social media is flooded with false information, ads, and other types of content, it is essential to recognise these fake accounts. From an online social media dataset, we offer a method for detecting fraudulent accounts. We employed boosting methods to improve the accuracy of the standard technique, rather than employing typical machine learning classifiers. By boosting weak learners, this method has resulted in a large improvement in accuracy. In this paper we will use accuracy comparison of Xgboost Classifier, Ada Boost Classifier and Gradient boosting Classifier.Xgboost performed brilliantly when compared with the previous work.

KEYWORDS:

machine learning, fake account detection,gradient boosting, accuracy

INTRODUCTION:

Social media is an essential part of everyone's life in today's modern world. The main aim of social media is to stay in contact with friends and share news, among other things. The number of people who use social media is rapidly growing. Instagram is a global social media platform that has recently grown in popularity. Instagram has over 1 billion active users, making it one of the most popular social media platforms. People with a large number of followers have been dubbed Social Media Influencers since the introduction of Instagram to the social media scene. These social media influencers have now become a popular place for businesses to promote their goods and services.

The widespread use of social media has turned out to be both a benefit and a liability for society. The use of social media for online fraud and the dissemination of false information is rapidly growing. On social media, fake accounts are the most popular source of false information. Businesses that spend a lot of money on social media influencers need to know if the following they've gotten is organic or not. As a result, there is a widespread demand for

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.

a fake account identification tool that can reliably determine whether or not an account is fake.

PROPOSED SOLUTION:

The gradient boosting algorithm is similar to the random forest algorithm in that it relies heavily on decision trees. We have modified the way we find fake accounts, using new approaches to locate them. Spam commenting, interaction rate, and artificial behaviour are some of the techniques used. The gradient boosting algorithm uses these inputs to build decision trees, which are then used in the gradient boosting algorithm. Even if some inputs are missing, this algorithm produces a result. This is the primary reason for using this algorithm. We were able to obtain extremely accurate results by using this algorithm. XGBoost performed brilliantly when compared with the previous work.It outperforms the accuracy of fake account identification by a large margin even with the default values of the hyper parameters provided in. Finally, we achieved a better result than earlier attempts.

GRADIENT BOOSTING MACHINE (GBM):

To produce final predictions, a Gradient Boosting Machine (GBM) combines predictions from multiple decision trees. Keep in mind that in a gradient boosting machine, all of the poor learners are decision trees.

But how is using a hundred decision trees better than using a single decision tree if we're using the same algorithm? What are the different ways that different decision trees capture different signals/information from data?

The trick is that each node in the decision tree uses a different subset of features to choose the best split. This means that the individual trees aren't all the same, and they can absorb different signals from the data as a result.

EXTREME GRADIENT BOOSTING MACHINE (XGBM):

Another widely used boosting algorithm is XGBoost (Extreme Gradient Boosting). In reality, XGBoost is just a tweaked GBM algorithm! XGBoost follows the same steps as GBM in terms of operation. XGBoost builds trees in a sequential fashion, attempting to fix previous trees errors.

However, there are a few features that render XGBoost slightly superior to GBM:

One of the most significant differences between XGBM and GBM is that XGBM uses parallel pre-processing (at the node level), making it faster.

Regularization strategies in XGBoost help to reduce overfitting and improve overall efficiency. Setting the XGBoost algorithm's hyperparameters allows you to choose the regularisation technique.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.

ALGORITHM:

GBM, XGBoost and AdaBoostclassifiers were trained and validated with training and validation sets after feature selectionand then accuracy was tested on the training set.

INPUT:

TrainData = The labeled training set (70%);

ValidationData= The validation dataset (10%)

TestData = Unlabeled dataset (20%)

OUTPUT:

Predictions = prediction from classifiers used.;

//ValidationData is used to validate the classifier predictions

1. Load TrainData

2. for all instances in TrainData

3. for each feature matrix fed to the CLASSIFIER [XGB, ADB, GBM]

4. train classifier

5. accuracy, precision, recall = PREDICTION.metrics

6. RESULT COMPARISON

METHODOLOGY:

1.Uploading the data.

2.Dataset pre-processing

3.Choosing a feature

4. A method for detecting fraudulent accounts and comparing the results.

Web Scraper:

A web scraper is an application that extracts information from a website. We gather relevant pieces of information from the social media site using Outwit hub, a Web scraper tool, when a user pastes a link to a social network account. Login activity, total likes, total comments, number of posts, number of followers, and number of followers are all data that we extract.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.

Calculation of Engagement rate:

An engagement rate is a metric that quantifies how much a post or storey has been shared on social media. It's the percentage of people who interact with a post. We can calculate the engagement rate by comparing the number of interactions to the number of followers. Likes, comments and shares are examples of interactions.Because the engagement rate is so easily computed, comparing popular and semi-popular accounts is pretty simple. Because lower audience involvement indicates that the account is phoney, this measure is one of the most important.

Engagement rate percentage= (total number of interactions)100

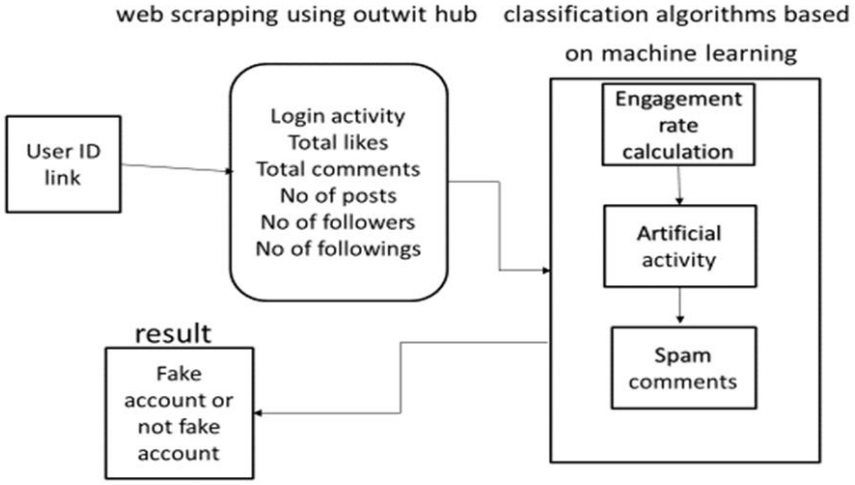
(Total number of followers)

Artificial Activity:

When the frequency of the above-mentioned behaviours is excessively high, normal social media behaviours like liking, commenting, and sharing become an artificial activity. The presence of activity also indicates that the account is being utilised by a bot. At this point, we look at how many likes, comments, and shares this account has received since it was created.If an account has a large number of likes or comments, it will be suspected as being false.We define huge as a quantity that is beyond the reach of the average social mediauser.

Spam Comments:

BOT comments are famous for being basic and lacking in creativity. At this point, the account's comments will be thoroughly examined. The total number of comments made by the user since the account was created will be compared to the average number of comments made by users in that particular OSN. If the disparity is significant, the account may be considered fraudulent.Commenting on links will result in the account beinglabelled as a fake.Comments of the same or similarnature will also be considered spam



Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.

EXPERIMENT AND RESULT:

1.Uploading the data.

2.Dataset pre-processing

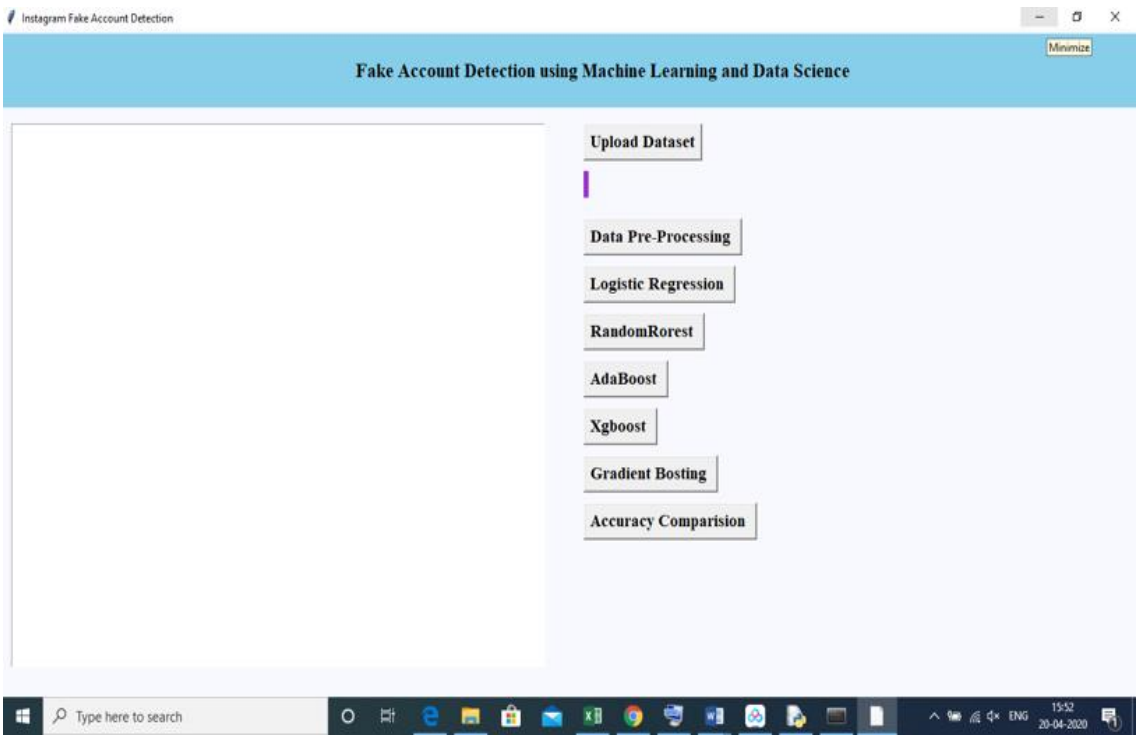
3.Choosing a feature

4. A method for detecting fraudulent accounts and comparing the results.

Table 1: Boosting Classifier Performance on chi2 Features

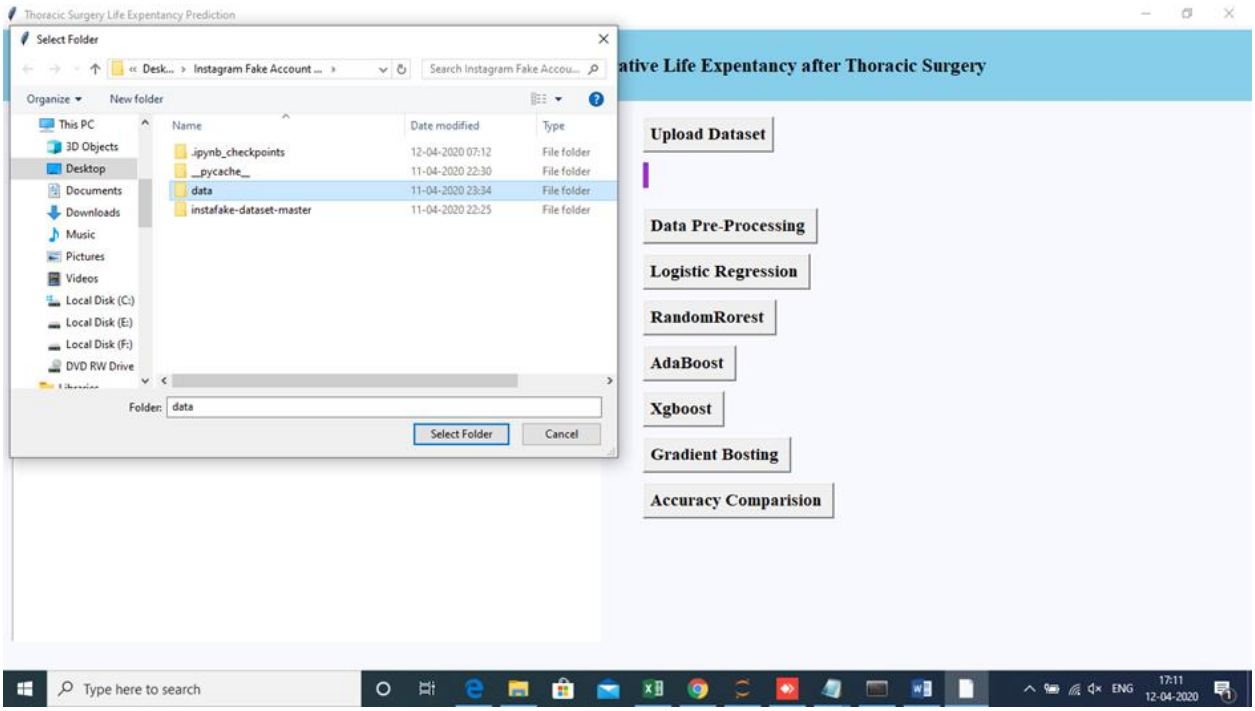
Classifiers	Feature selection	Xgboost	Adaboost	GBM
Accuracy	Chi2	0.958	0.942	0.952
Precision		0.951	0.911	0.939
Recall		0.898	0.887	0.906

Boosting classifiers outperformed typical machine learning classifiers by a significant margin. The default parameter values for these boosting classifiers were used. XGBoost obtained the value of 95 percent, which is slightly higher than other chi2 enabled products.



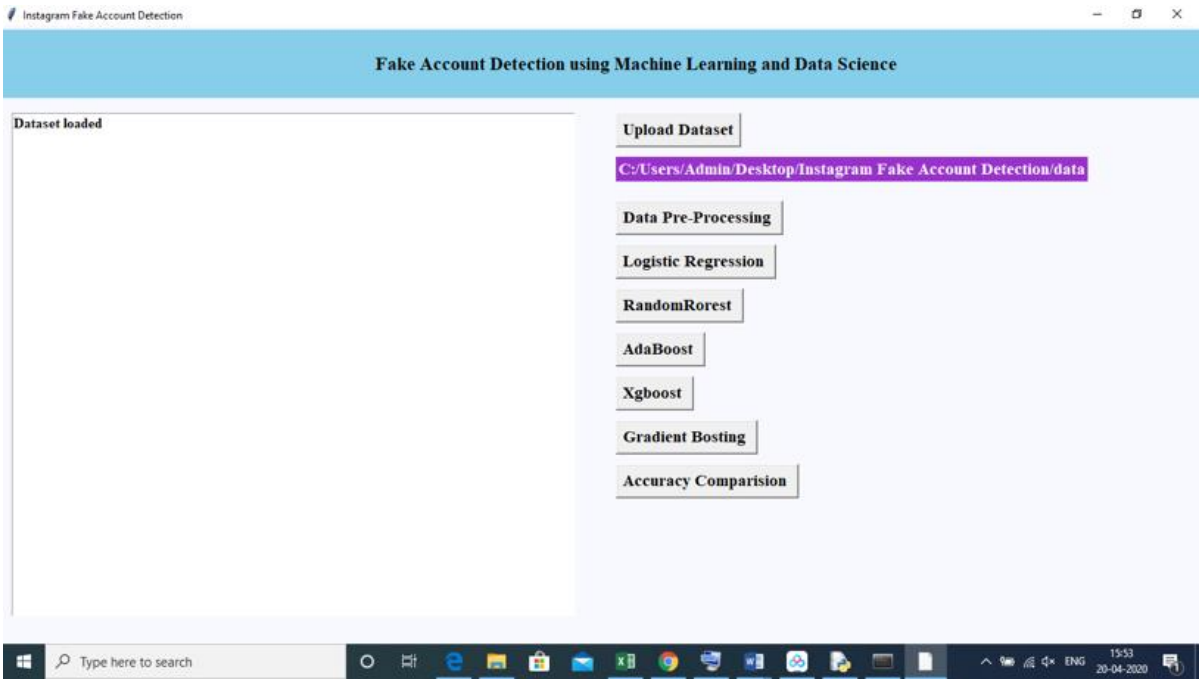
After above screen will be opened and Need select the dataset directory by clicking on upload button.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.

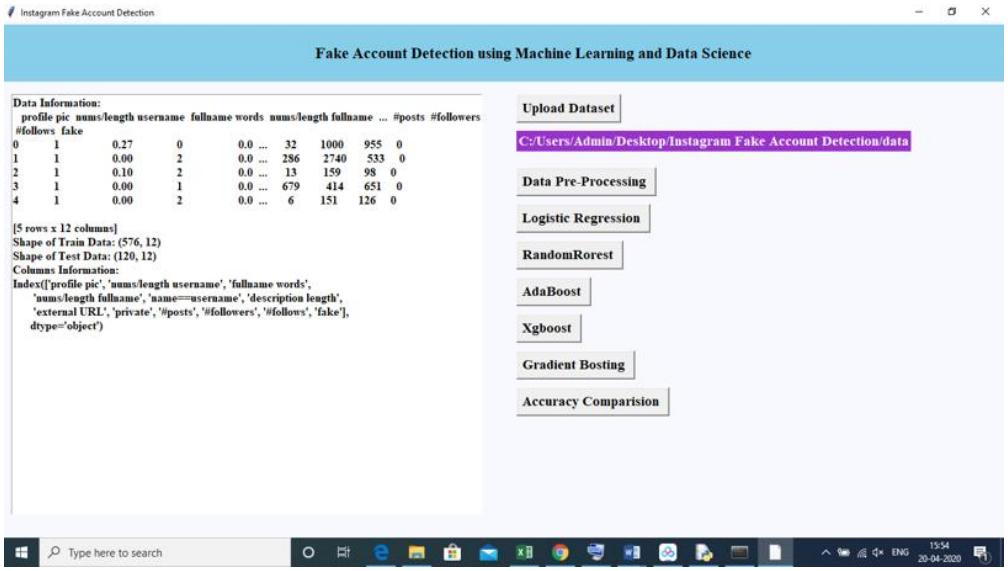


Data is uploaded.

Now click on pre-process button

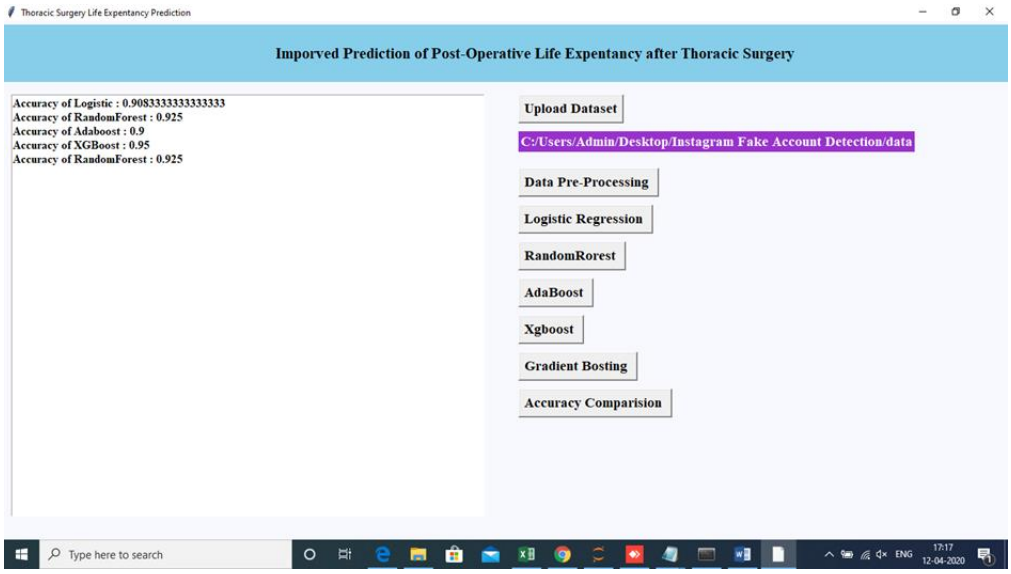


Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.



Data Pre-processing will be done and will show the data information

Now click on Logistic Regression, Xgboost, Adaboost, Gradient boosting buttons

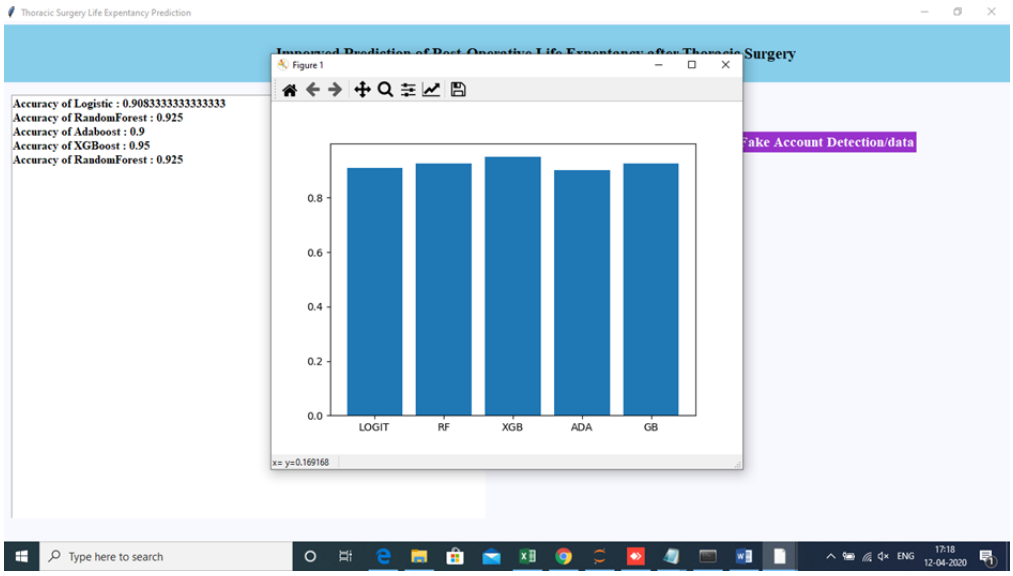


Will run the logistic algorithm, xgboost, adaboost, gradient boosting algorithm on the given data and will give accuracy on Test data.

GBM AND XGBOOST ACCURACY:

Now click on “Accuracy Comparison”.From the algorithm we can tell the Xgboost performed well.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.



CONCLUSION:

We said before that there was a lack of a gold standard public dataset for analysis, thus we had to use active learning. Manually labelling the data may improve the model's performance. The use of Extreme Gradient Boosting to detect fraudulent accounts is still relatively new and on the rise. There are numerous branches to investigate.As previously stated, we did not perform deep hyperparameter tuning in our suggested strategy or trials. Tuning hyperparameters is both expensive and time-consuming. Finding the optimal collection of parameters might be difficult. XGBoost, on the other hand, fared better with default values, achieving accuracy of up to 95%.

FUTURE SCOPE:

Finally, we intend to enrich the dataset further and look forward to observing the results of other elements of the boosting methods.

REFERENCES:

1. “Detection of Fake Twitter accounts with Machine Learning Algorithms” Ilhan Aydin, Mehmet sevi, Mehmet Umut salurJanuary 2019.
2. “Detecting Fake accounts on Social Media” Sarah Khaled, Neamat el Tazi, Hoda M.O. MokhtarJanuary 2019.
3. “Detection of fake profile in online social networks using Machine Learning” Naman Singh, Tushar Sharma, Abha Thakral, Tanupriya ChoudhuryAugust 2018
4. “Twitter fake account detection”, Buket Ersahin, Ozlem Aktas, Deniz kilinc, Ceyhun AkyolNovember 2017.
5. “A new heuristic of the decision tree induction” Ning li, li Zhao, ai-Xia chen, Ging-Wu meng, Guo-fang ZhangAugust 2009.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 6857 - 6865
Received 25 April 2021; Accepted 08 May 2021.

6. “statistical machine learning used in integrated anti-spam system” Peng-Fei Zhang, Yu-Jie su, Cong wangOctober 2007.
7. “A study and application on machine learning of artificial Intelligence” ming Xue, Changjun zhuJuly 2009.
8. “Learning-based road crack detection using gradient boost decision tree” Peng sheng, li chen, Jing tianJune 2018.
9. “Verifying the value and veracity of extreme gradient boosted decision trees on a variety of datasets” Aditya Gupta, Kunal Gusain, Bhavya popliJanuary 2018.