2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) | 979-8-3503-1156-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/WIECON-ECE57977.2022.10150753

FAKE PROFILE IDENTIFICATION USING MACHINE LEARNING

T.Sudhakar
School of Computer Science and Engineering
VIT-AP University, Vijayawada
Andrapradesh
Email: tsudhakar105@gmail.com

Bhuvana Chendrica Gogineni
School of Computer Science and Engineering
Ira A. Fulton Schools of Engineering Arizona State University
United States
Email: goginenibhuvana11@gmail.com

J. Vijaya
Department Data Science and Artificial Intelligence
International Institute of Information Technology-Naya Raipur
chhattisgarh
Email: vijayacsedept@gmail.com

***Abstract*—Online social networks have permeated our social lives in the current generation. These sites have allowed us to see our social lives differently than they did in the past. Nowadays we can connect with new friends and maintain relationships with them via social and personal activities become quite easy. Online Social Networks (OSN) are contributed in all areas such as Research in all domains, Job-related areas, Technology oriented areas, Health care, and business-oriented areas, Information gathering and data collection, and so on. One of the biggest problems on these social media platforms is fake profiles. Impersonating to be someone else and causing harm and defamation to the real person or advertising or popularizing removed propaganda on someone's name to get more benefit is the motto of such profile creators. There have been many studies regarding these fake accounts and how can they be mitigated. Many approaches such as graph-level activities or feature analysis have been taken into consideration to identify fake profiles. These methods are outdated when compared to arising issues of these days. In this paper, we proposed a technique using machine learning for fake profile detection which is efficient. . The benchmark data set is collected and mixed with manual data first furthermore; a data cleaning technique is used to present the data more feasibly. Then the preprocessed data is used for model building with sufficient information such as profile name, profile ID name, number of followers, and so on. We added Cross validation process where many training algorithms are implemented on the given data and are then tested on the same data. Based on the experiments the RF classifier performed better than the other classification methods. The Random Forest classifier is used to forecast the profile whether is fake or genuine in an efficient way.**

***Index Terms*—Fake profile identification; Machine learning; Classification; Random Forest**

I. Introduction

On the internet, you can access a variety of opportunities and connections. You're probably familiar with popular social networking sites like Facebook, WhatsApp, Instagram, and Snapchat. In addition to these forms of social interaction, our current generation also participates in many other forms of interaction [1-2]. Social networking sites are very easy for teachers to train youngsters and Teachers in this era have become very familiar with these websites, giving online lectures, giving assignments, holding discussions, and so on, which improves education significantly. Using these social networking sites, employers can hire human beings who are skilled and enthusiastic about the work; their history can easily be looked into by utilizing these websites. These platforms are free however some cost the membership price and make use of this for commercial enterprise functions and the rest of them elevate cash with the aid of the usage of advertising [3-4]. However, it also has some downsides and one of those is fake profiles. They are usually a result of the simple lack of engagement with people face-to-faces and this can often lead to invitations that we wouldn't normally receive if these fake profiles weren't present on social networks [5]. Due to the wide use of social networks, there have been many studies done in this domain. Among them, Devakunchari. (2018) have done a survey and their findings show that the percentage of internet users who have been impersonated online is 82%. Also, the percentage of those who have been deceived is 9% and 22% have been tricked into giving out personal information. There have been many studies to identify fake accounts in the OSN's platform that are inaccurate and can be only useful at the time of naive attacks as they can be invaded easily. There is no feasible solution that could be 100% accurate to control this problem [6].

A. Problem Statement

Many issues, including fake profiles and online impersonation, have arisen in today's online social networks. No one has developed a potential solution for these issues as of yet. To safeguard human social lives, we want to provide a novel model for computerized fake profile early detection in this project. We can also make it simpler for websites to change the large diversity of profiles by using our automatic detecting technology which is hard to accomplish manually.

II. LITERATURE SURVEY

Many issues, including fake profiles and online impersonation, have arisen in today's online social networks. No one has developed a potential solution for these issues as of yet. To safeguard human social lives, we want to provide a novel model for computerized fake profile early detection in this project. We can also make it simpler for websites to change the large diversity of profiles by using our automatic detecting technology which is hard to accomplish manually.

To discover the traits or a mixture of them that help to differentiate between real and false records, several fake document focus approaches rely on the examination of a person's interpersonal organization profiles. Specifically, a classifier that can recognize bogus data is built using machine learning techniques after many attributes are obtained from the profiles and posts.

Padmavati et al. approach the problem of fake accounts on social media by a method using Deterministic Finite Automata (DFA) [7]. The paper analyzes the features of the existing user and their friends by creating an accounting pattern. The pattern is made with regular expressions based on some attributes such as the working and living community and so on which are used for pattern matching with any friend requests. The drawback of this approach is that the generation of regular expression takes quite a long for a person who has friends in many communities. The authors contend that the approach could be even more efficient for the real world.

In their article, Mohammadreza et al. employed graph analysis and classification algorithms to examine the issue of phony accounts on social networks. Twitter was the social media platform of choice. They developed a strategy based on how similar the user's friends were. Before extracting new features using Principal Component Analysis (PCA), it first employs the buddy similarity criterion from the network graph [8]. Next, the data is balanced and delivered to the classifier using the synthetic minority oversampling method (SMOTE). A medium Gaussian SVM classifier was selected after utilizing the cross-validation procedure since it has an AUC of 1. This method's flaw is that phony accounts must only function within the network to avoid being detected by looking at the accounts of their friends. The authors assert that in the future, a new technique would be introduced that could determine if an account is real or fraudulent at the moment of registration or even before any user activity on the network.

Srinivas Rao et al. have tried machine learning and Natural Language Processing (NLP) in their paper for fake profile detection [9]. The authors used Facebook profiles as their dataset. The process has three phases namely NLP preprocessing, PCA, and learning algorithms. They pre-processed the data using tokenization, stop word removal, Stemming and Lemmatization. PCA is done to extract the fundamental values from the table. Later, two ML algorithms named SVM NB are used to classify profiles. The observation after evaluation of their approach showed that the detection accuracy improved when these algorithms were used.

Stringhini examined the marketplaces for Twitter supporters. They list the characteristics of Twitter aficionado adverts and group the patrons of the business sectors. According to the authors, two main categories of bills chase the "client": hacked accounts and fake accounts (often known as "sybils"), whose suppliers do not assume that the number of their followers is growing [10].

Clients of adherent marketplaces may include well-known individuals or politicians who want to give the appearance of having a larger fan following, or they may include cyber criminals who want to make their files appear consistently real so they may erratically spread malware and spam. Thomas examines the counterfeit money used to send out Twitter spam. Based on emotions like happiness, sadness, rage, fear, etc., Nancy Agarwal et al. classified the users as phony or real. They test it out using Facebook users' postings. Twelve emotion-based characteristics are used to train the detection model [11]. The author's study is based on the observation that actual users post with a range of emotions, but fake users, who are assigned to certain professions, post with a consistent set of feelings. Additionally, noise reduction is performed. Finally, the detection model has been trained using machine learning techniques including NB, JRip, SVM, and RF.

Machine learning techniques were employed by Ananya et al. in 2021 ICRITO to identify phony social media profiles [12]. They received the data from Kaggle, an open-source website where data sets are stored for public use. The information came from Weibo, a Chinese version of Twitter and a popular networking site. Later, they used five supervised learning models for training and to check which one gave a better test score (cross-validation). Out of all 5 techniques they chose Gradient Boosting Classifier and Random Forest Classifier as they showed a better performance concerning others. Finally, after which they chose a random forest classifier as it gave a 1% better result than the gradient boosting classifier. They aim to make an automated system that can train and take more attributes than the limited ones in this paper for their future work.

Preethi Harris et al.described that fake Instagram profiles could be identified using machine learning [13]. The profile data from social media called Instagram was taken from the Kaggle website. They employed classification algorithms such as SVM, KNN, RF, NB, and XG Boost to train the model. After computing the accuracy and confusion matrix, the RF classifier stood out as the suitable model for the data set with the best results of prediction. Later the IDs of the Fake profiles are written into a data dictionary.

In 2018, Abhishek Narayanan et al. discussed recognizing fake profiles in their paper and their data set was taken from the social media named Twitter. At first feature extraction on the data was done on which machine learning algorithms namely SVM, RF, and LR performed which gave a very much appreciated result for the random forest in the end. Later after performing some accuracy testing and confusion matrices random forest classifier stood out with 88% of precise prediction of fake profiles on Twitter.

It was more efficient and comparatively took the shortest to achieve the results. Their future work is intended towards ensuring the security of the users while surfing other using social media[14].

In 2012, Mauro Conti et al. discussed through the paper the ways the issue can tackle [15]. The first thing done was to check if a particular profile is similar to the population of real users. Then they used graph structures for fake profile detection. They observed the user's connection that is the friends list whether there is a greater number of random ones or whether there are certain numbers of mutual friends. Social network structural analysis is done to check for a fake profile. In their future work, they would intend to come up with a better mechanism and extend the characterization to online interactions such as tags, friendship requests, and rate of acceptance of requests. etc.

III. Proposed Work

The suggested framework below outlines the steps that must be taken in order to detect fraudulent profiles, with active learning taking place as a consequence of feedback from the classification algorithm's results, which may also be shown in the system model diagram (Figure 1). The steps in the procedure are as follows:

- Data collection and cleaning of data
- Cross Validation is performed to check the perfect model for the data
- The chosen model is trained with the data set
- Then pipe-lining is performed for even better accuracy
- The model is then evaluated with a test data set

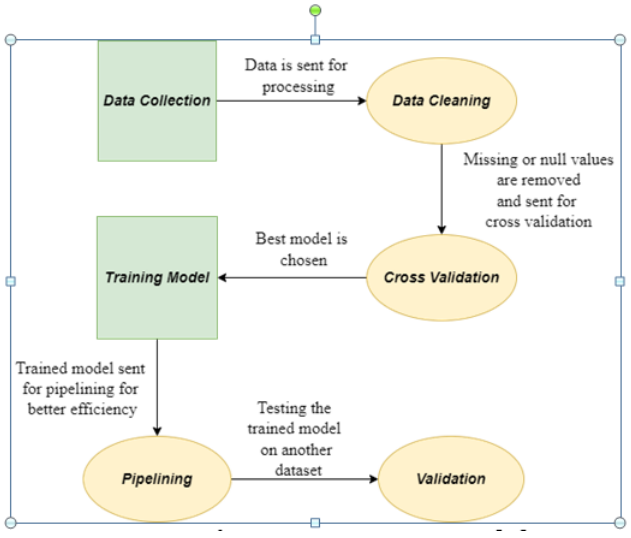

Fig. 1. Proposed System Model

A. Data set

The data set is accumulated from two primary sources. The first one is from Kaggle [16] and the other one is from Git hub [17].

Both are open-source platforms where data sets are kept for public usage. we have used common attributes from both data sets as any social media would have such as user name, number of followers, number of following, and so on. This was done to see if better accuracy of prediction could be achieved with more data.The data set is manually added with 200-row entries and is further used in the process. The data sets used are the already existing ones and the one that was manually combined with the already existing ones. In this way, we can use the data set for prediction and also find out if more amounts are useful in this case

We need a data-set that contains both fake and genuine profiles. Data-set is divided into 2 training and testing data. The classification algorithm uses the training data for training and the testing data set is used to check the efficiency of the model produced. Two data sets are taken to check the accuracy of the model when trained with more amount of data. One data set(data set-1) has 556 entries of data while the second data set(data set-2) has 776 entries of data. The first data set is manually combined with 200 entries of data from another data set to form the second data set. Attributes taken from the profiles that are considered in the process of training for the identification of fake profile identification are as follows [Figure 2]:

Attributes	Description
Profile Picture	User has a profile picture or not
Full name words	Number of words in tokens
Bio/Description length	Description length in characters
External URL	Has external URL or not
Private	Private account or not
Posts	Number of posts
Followers	Number of followers
Follows	Number of follows

Fig. 2. List of attributes and description

B. Cross-validation

Cross-validation is a process where many training algorithms are implemented on the given data and are then tested on the same data. The mean scores are then displayed; the scores with higher values represent a better ability to give better output with the data. As every data set has its characteristics, so each data set has different optimal algorithms to train which could be found with the help of cross-validation.It is a very simple yet efficient way of making better use of the already present model.

C. *Classification Algorithms*

In this project, there are many topics and algorithms involved. A briefing of such topics can be seen in this chapter to get a better understanding of the process during the implementation. Generally, there are 2 categories named regression and classification in machine learning. These are used according to the data set and the type of output that the user requires. Classification is used when the data set is set and limited. It is used when the output required is in the format of yes or no, true or false type. Meanwhile, regression is used when the data is continuous and it is quite often used in weather predictions. In this project, four machine learning algorithms are used named Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression, and Gaussian Naive Bayes.

1) ***Random Forest****:* It is a supervised ensemble learning technique that constructs many decision trees during the training phase, then uses a mean voting mechanism to choose the top decision trees for prediction. The data is split randomly into many data samples which are then split into trees that are trained and tested for prediction. The prediction score is then finalized by the voting scheme in between the trees.

2) ***Gradient Boosting*** *:* It is a classification algorithm that uses an additive predictive model by combining various decision trees. It is similar to the random forest model but it relies on the intuition that the best possible next model would be the one that gives a minimum number of errors when combined with the previous one. It is a combination of many weak learning models to create a strong prediction model

3) ***Logistic Regression*** *:* Both a predictive analysis technique and the idea of probability are used. This is a statistical method for investigating data when the outcome is controlled by one or more independent factors. Evaluation-based logistic regression selects parameters that increase the likelihood of finding the case values. It generates the formula coefficients to predict a legitimate transformation of the obligation of reality of the aspect of interest.

4) ***Gaussian Naive Bayes*** *:* Additionally, it uses supervised machine learning. It is also a particular use of the Naive Bayes method where the characteristics have continuous values. This algorithm assumes that all the features follow a Gaussian distribution which is also called a normal distribution. This model fits simply by using mean and standard deviation

D. *Chosen Model- Random Forest*

We utilised chosen Random Forest Classifier from the cross validation process in this case. It is a supervised learning algorithm that performs both classification and regression. Classification is used when there is a requirement for prediction or classification with a certain fixed amount of data whereas Regression is used when there is a requirement for prediction for continuous data such as in the stock market.Random Forest is an algorithm that makes a forest that by selecting the best decision trees employing voting. It divides the data set into subsets to make several decision trees. So, the larger the data the more accurate the outcome. The Random Forest Classifier's process is as follows:

- Selection of random data samples from the data set
- Construction of a decision tree for every sample which then gives the predicted result from each tree
- Voting for every result is done
- The selection of most voted predicted result is chosen as the final prediction result

E. *Pipe-lining Process*

Pipe-lining is iterative as each step is repeated numerous times to improve the accuracy of the model and to create a successful algorithm. Pipe-lining is performed on the model and this technique is used to compare and analyze data that contain similar qualities or are in a linear sequence of data transforms to be chained together culminating in a modeling process that can be evaluated. This splits the data into reusable independent parts which are combined again to form a model. It operates by correlating the data together in a model that can be tested and evaluated to achieve an outcome that increases the efficiency of the model. It ensures that the data used for preparations that are performed can be reusable. It automates the workflow it takes to generate a machine-learning model that is efficient yet accurate and precise to the best.

F. *Evaluation Matrices*

This project used different types of plotting techniques such as correlation and confusion matrices for a better understanding of the subject. The correlation graph is used when there is a requirement of checking for the reliability of the attributes. A confusion matrix is a plotting matrix that shows a better pictorial representation of an output that could be understood quite easily.

IV. Implementation and Result Analysis

A. *Implementation*

The implementation process is given below

1) Load, Inspect and Clean data
   - Check correlation(the feature comparison done to check the relativity of the attributes in the data set) between attributes. Any data set can be chosen if its correlation heat map of it is approximately more than zero. The correlation graph was approximately above 0, so the data sets are favorable to be trained into the model
   - Check for missing or null values. Both data sets show no null values or miss placed data types, so there is no need for any furnishing of the data needed, and can be processed for further steps.
2) Creating percentage share of Train and Test data set. From the training data set-1, we have used 60% of the data for training and 40% for testing to check the model before implementing as this particular data set showed more positive results than other percentage shares. On the other hand, from the training data set-2, we have used 75% of the data for training and 25% for testing to check the model

3) Cross Validation is done to check which model is the best fit for the data set. The models considered for testing are Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression, and Gaussian NB. It can be seen in the Table 1 and 2 that Random Forest Classifier is showing the best validation score for both the data sets followed by Gradient Boosting Classifier. So Random Forest Classifier model is chosen for training

TABLE I
CROSS VALIDATION RESULTS OF DATA SET 1

MODELS	Training score	Validation score
Random Forest	1.000	0.919
Gradient Boosting	1.000	0.913
Logistic Regression	0.891	0.867
Gaussian Naive Bayes	0.698	0.690

TABLE II
CROSS VALIDATION RESULTS OF DATA SET 2

MODELS	Training score	Validation score
Random Forest	1.000	0.929
Gradient Boosting	0.997	0.926
Logistic Regression	0.921	0.916
Gaussian Naive Bayes	0.753	0.747

4) Applying Random Forest Classifier using grid search by passing the parameters such as the maximum number of trees(n-estimators) and the maximum depth of the tree (max depth). The best parameters shown from the process are chosen after the training is done and the mean training and test scores it got with those parameters
5) Then Pipe-lining is done. The test scores after pipe-lining the model brought quite a difference of 4% more accuracy in the model of data set-2(94% score) than the model of data set-1(90% score)
6) Then test the model by passing the test data set to the trained model is done and a confusion matrix is drawn to have a clear view of the results

B. Result Analysis

From Table 3 and 4, the confusion matrix shows that only 6 profiles were detected wrong for the model built on data set-2 whereas 8 profiles were detected wrong by the model built based on data set-1. This shows that the model trained by data set-2 is giving more efficient results compared to the model created by data set-1 which in return explains that more the data given to a model more precision and accuracy will always be achieved.

The below numerical values show the classification report details done on the test data in a detailed manner showing the percentage of preciseness it achieved when both the models are tested on the same test data set of 120 entries. Table 5 an 6 show a clear boundary of percentages of prediction where the model of data set-2 is seen to predict more accurately (95%) than the other one (93%).

TABLE III
CROSS VALIDATION RESULTS OF DATA SET 1 FOR TEXT(33.0,0.5,ACTUAL VALUES)

		Predicting fake account		
		Genuine	Fake	Total
Actual values	Genuine	56	4	60
	Fake	4	56	60
	Total	60	60	120

TABLE IV
CROSS VALIDATION RESULTS OF DATA SET 2 FOR TEXT(33.0,0.5,ACTUAL VALUES)

		Predicting fake account		
		Genuine	Fake	Total
Actual values	Genuine	58	2	60
	Fake	4	56	60
	Total	62	58	120

And also the random forest classifier algorithm could efficiently detect 93% of genuine and 93% of fake accounts for the model created with the data set-1. The random forest classifier algorithm could efficiently detect 94% of genuine and 97% of fake accounts for the model created with the data set-2.

TABLE V
FINAL EVALUATION SCORE OF MODELS OF DATA SETS 1

-	Precision	Recall	F1-Score	Support
Genuine	0.93	0.93	0.93	60
Fake	0.93	0.93	0.93	60
Accuracy	-	-	0.93	120
Macro average	0.93	0.93	0.93	120
Weighted average	0.93	0.93	0.93	120

TABLE VI
FINAL EVALUATION SCORE OF MODELS OF DATA SETS 2

-	Precision	Recall	F1-Score	Support
Genuine	0.94	0.97	0.95	60
Fake	0.97	0.93	0.95	60
Accuracy	-	-	0.95	120
Macro average	0.95	0.95	0.95	120
Weighted average	0.95	0.95	0.95	120

C. Result comparison with Base classifiers

Table 6 represent the classification accuracy of the proposed model compared with the base classifiers such as DT, KNN, NB, SVM, and ANN for the created data set. It shows that the proposed model performs better than the existing base clasifier models in fake profile detection system.

D. Result comparison with Existing models

Table 7 represent the classification accuracy of the proposed model compared with three proposed work by the authors[18-20] for the created data set. It shows that the proposed model performs better than the existing classifier models in fake profile detection system.

TABLE VII
RESULT COMPARISON WITH BASE CLASSIFIERS

MODELS	P.RF	DT	KNN	SVM	NB	ANN
Accuracy(%)	95	88	87	88.5	77	83

Adhikari and Dutta [18] provide recognizably false LinkedIn accounts as evidence. The study shows that using restricted profile records as input, fake profiles may be identified with 84% accuracy and 2.44% false negatives Techniques including principal thing evaluation, neural networks, and SVMs are used. Highlights include, among other things, the variety of languages spoken, education, skills, recommendations, hobbies, and honors. The characteristics of profiles that have been identified as false and uploaded on exotic websites are used as a starting point.

The Chu strives to distinguish between Twitter debts run by people, bots, or cyborgs (i.e., humans and bots working together) [19]. An Orthogonal Sparse Bigram text content classifier that employs pairs of words as features is used to identify spamming archives as part of the detection problem formulation.

In his work, Nazir discusses how to spot and identify fraudulent profiles in social media-based online gaming apps. The study examines a Facebook application called "Fighters club," an online game that is said to offer rewards and game play advantages to users who encourage their friends to play [20]. According to the authors, by providing such incentives, the sport encourages its players to create false profiles e-user would increase a motivating pressure of an incentive for himself/herself by giving these fake profiles into the game.

TABLE VIII
RESULT COMPARISON WITH EXISTING MODELS

MODELS	P.RF	Dutta[18]	Chu[19]	Nazir[20]
Accuracy(%)	95	84	92	90.5

V. CONCLUSION

A comparison is done between 2 models trained with different data sets where one had more amount of data than the other. The results turned out better for the data set with more data in it. We have given a framework with the use of which we can become aware of fake profiles in any online social community via the usage of the Random Forest Classifier model and pipe-lining with a very excessive efficiency of as high as 95% on average. In the future, we wish to classify profiles by taking a larger amount of data with different data types. Even use some data preprocessing methods to make use of efficient data when there are larger data sets. Also, to come up with a system that can identify a fake profile by giving required attribute inputs to the mode

REFERENCES

[1] Prabhu Kavin, B., et al. "Machine learning-based secure data acquisition for fake accounts detection in future mobile communication networks." Wireless Communications and Mobile Computing 2022 (2022).

[2] Salman, Fatima Maher, and Samy S. Abu-Naser. "Classification of Real and Fake Human Faces Using Deep Learning." International Journal of Academic Engineering Research (IJAER) 6.3 (2022).

[3] Kenny, Ryan, et al. "Duped by bots: why some are better than others at detecting fake social media personas." Human factors (2022): 00187208211072642.

[4] Purba, Kristo Radion, David Asirvatham, and Raja Kumar Murugesan. "Classification of instagram fake users using supervised machine learning algorithms." International Journal of Electrical and Computer Engineering 10.3 (2020): 2763.

[5] Samala Durga Reddy. "Fake Profile Identification using Machine Learning" In Dec 2019 IRJET journal Volume: 06 Issue:12 pp.1145-1150

[6] Devakunchari Ramalingam, Valliyammai Chinnaiah. "Fake profile detection techniques in large-scale online social networks:A comprehensive review"In 2018 Computers Electrical Engineering,Volume 65,Pages 165-177

[7] Padmaveni Krishnan D.John Aravindhar Palagati Bhanu Prakash Reddy "Finite Automata for Fake Profile Identification in Online Social Networks." Proc. Of ICICCS 2020, Part Number: CFP20K74-ART

[8] Mohammadreza Mohammadrezaei, Mohammad Ebrahim Shiri and Amir Masoud Rahmani "Identifying Fake Accounts on Social Networks Based on Graph Analysis and Classification Algorithms" Security and Communication Networks, Volume 2018, Article ID 5923156, 8 pages

[9] P. Srinivas Rao, Dr. Jayadev Gyani, Dr. G. Narasimha "Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4133-4136

[10] Stringhini, Gianluca, Gang Wang, Manuel Egele, Christopher Kruegel, Giovanni Vigna, Haitao Zheng, and Ben Y. Zhao. "Follow the green: growth and dynamics in twitter follower markets." In Proceedings of the 2013 conference on Internet measurement conference, pp. 163-176. ACM, 2013

[11] Mudasir Ahmad wani, Nancy Agarwal, Suraiya Jabin and Syed Zeeshan Hussain. "Analyzing Real and Fake users in Facebook Network based on Emotions." In the proceedings of the 2019 11th International Conference on Communication Systems Networks (COMSNETS), pp. 110-117.

[12] Ananya Bhattacharya, Ruchika Bathla, Ajay Rana, Ginni Arora." Application of Machine Learning Techniques in Detecting Fake Profiles on Social Media"In the 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. Sep 3-4, 2021

[13] Preethi Harris, Gojal J, Chitra R, Anithra S."Fake Instagram Profile Identification and Classification using Machine Learning". In 2021 2nd Global Conference for Advancement in Technology (GCAT) Bangalore, India. Oct 1-3, 2021

[14] Abhishek Narayanan, Anmol Garg, Isha Arora, Tulika Sureka, Manjula Sridhar, Prasad H B."IronSense: Towards the Identification of Fake User-Profiles on Twitter Using Machine Learning"In 2018 Fourteenth International Conference on Information Processing (ICINPRO)

[15] Mauro Conti, Radha Poovendran, Marco Secchiero. "FakeBook: Detecting Fake Profiles in On-line Social Networks"In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,pp.1071-1078

[16] https://www.kaggle.com/datasets/bitandatom/social-network-fake-account-dataset

[17] https://https://github.com/harshitkgupta/Fake-Profile-Detection-using-ML

[18] Haq, Amin Ul, et al. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." Mobile Information Systems, 2018.Adikari, Shalinda, and Kaushik Dutta. "Identifying Fake Profiles in LinkedIn." In PACIS, p. 278. 2015

[19] Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. "Who is tweeting on Twitter: human, bot, or cyborg?" Proc. Of the 26th annual Computer security applications conference 2010, pp.21-30

[20] Nazir, Atif, Saqib Raza, Chen-Nee Chuah, Burkhard Schipper, and C. A. Davis. "Ghostbusting Facebook: Detecting and Characterizing Phantom Profiles in Online Social Gaming Applications." In WOSN. 2010