IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Evaluating and Detecting Fake Users in Social Media by Random Forest

Mahi Maanas Reddy[1], Shruti Sridhar[2], V. Maria Anu[3], Dr Punitha K[4]
UG Students, School of Computer Science and Engineering[1,2]
Associate Professor, School of Computer Science and Engineering[3,4]
Vellore Institute of Technology, Chennai, India

Abstract: *Currently, users have been engaging in conversations, sharing information and producing web content via social media platforms. But in recent times, many users have been using these platforms to conduct identity faults, payment frauds, and many more without the knowledge of the actual user. For example: - On Instagram, according to the latest analysis, there are around 95 million fake accounts compared to the total number of users, which amount to 1 billion. Therefore, there are nearly 10% of fake accounts active at present. The obtained dataset lies approximately in thousands. Hence, we used GANs and deep learning to broaden the data to around 1 lakh. The conventional methods used for distinguishing between real and fake accounts were ineffective. Adopting machine learning-based approaches allowed us to identify fake accounts that can mislead users. The dataset is pre-processed using several Python tools, and a comparison model is created to identify a practical solution appropriate for the dataset that has been provided.*

Keywords: Logistic Regression, Random Forest, XGBoost, Support Vector Machine, Generative adversarial networks.

I. INTRODUCTION

A social network is a website or internet application that promotes social interaction by helping users connect with people who share their interests, create forums for conversation, and exchange information, e.g.:- Twitter, Facebook, Instagram etc.

However, users are increasing daily, and many accounts created are deceptive. Online accounts that don't belong to genuine individuals are known as fake accounts. Some are made for amusement, some for fraud, and some for disseminating false information.

One of their most popular uses is to falsely increase followers. The users who want influencers or businesses looking for rapid expansion can get thousands of Instagram followers, and it's all automated. The various drawbacks of fake profiles on any social media platform are:

- **Decrease engagement rate** - The ratio of likes you receive to followers you have is used to compute the engagement rate. In the long run, these fake accounts will stop engaging with your posts, decreasing the engagement rate and reducing your profile's general visibility.
- **Damage your prestige** - Many customers consider the number of likes, comments and followers before purchasing from a small business on any social media platform. They will begin to lose faith in your brand if they notice that your following list consists of fraudulent identities and that your posts are sprinkled with generic spam comments.
- **Alter Your Metrics** - When correctly applied, analytics may show you what you're doing right and wrong and point out areas for improvement. But, having excessive phoney followers can make it difficult to examine your data. How can you determine your engagement, conversion, or click-through rates if you don't know your true follower count?
- **May lead to the ban of your accounts** - The creation of fake accounts cannot be stopped; instead, they are dismissed. Additionally, if moderators suspect you have too many bogus followers or are using other prohibited tools, your account can be restricted, suspended, or possibly banned.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

In this paper, we have used machine learning to help predict the fakeness of a social media account. Machine learning algorithms employ computational techniques to learn information directly from data without using a preexisting equation as a model. As the number of samples available for learning rises, the algorithms adaptively improve their performance.

We have used additional attributes besides the general dataset to increase accuracy. They include profile picture, username length, description and description length, external URL and profile security (public or private).

The data gathered is then utilized to generate more data, leading to more combinations of data scenarios that, when put into action, can produce more accurate results. The GANs algorithm creates more information from the current data set. Further, we verified the accuracy of our classifiers with Random Forest, Logistic Regression, SVM and XGBoost.

The research based on the previous work on fake users, GANs and Machine learning algorithms is described in Section 2. The generated dataset and the working of the code are discussed, and the generated data is then pre-processed and used to classify fake and real accounts in Section 3. The working of the most accurate algorithm is shown in Section 4, followed by results and summary in Sections 5 & 6 respectively. The conclusion is presented in Section 7.

II. LITERATURE SURVEY

The source codes of the most popular opensource Instagram bots were studied to collect data [1]. The generator creates false data from the distribution of accurate data. Fake data is sent to a discriminator, which will determine whether they are factual data [16].

Several traditional and neural network-based learning methods have been used as classifiers. Naïve Bayes, logistic regression and SVM are used [1]. A feature-based data mining technique is suggested in this study. Decision trees automatically separate bogus users from genuine users. Their accuracy varies depending on the algorithm, from 70.3% to 92.1% [2].

The predictions show that the neural network algorithm obtained 93% accuracy. Using natural language processing methods for skin detection will improve accuracy [7]. According to the finding, SVM-NN has a classification accuracy of 98% and has achieved extraordinary precision compared to the other two classifiers. SVM with polynomial kernel and PCA selected features provides the highest accuracy and lowest false negative percentage [13].

To detect fake accounts, Sybil Rank is used. It is a practical and efficient fake account inference technique that enables OSNs to rank accounts based on their apparent fakeness 5]. Sybil Walk maintains the benefits of current random walk-based approaches while overcoming drawbacks [9]. The SPP programme is introduced to safeguard Facebook user privacy better. The SPP software can be separated into three layers of security by this design [8].

An alternate strategy that can be employed as a leading or supplementary detection approach is nonverbal behavior monitoring for deception detection. Testing these systems across many platforms is necessary to progress the study of social media identity deception detection [10]. The paper examines how the fundamental ideas of social honeypots, generating spam profiles, and adaptive and constant spam detection can efficiently detect spam profiles [11].

The adaptive detection approach significantly increases the detection and reaches accuracy on par with the paired record comparison algorithm [12]. The inference is carried out in the simplest version of GANG using Loopy Belief Propagation [14]. This unique, innovative functionality, called Circle, can create circles for many categories of friends, including friends who share a hobby, relatives, and classmates [15].

The study has done away with the necessity for manual prediction of a fake account [4]. The Evasion strategies look at four cuttingedge alternatives based on the analysis; adopting the novel feature set dramatically increases the detection rate [6].

This paper demonstrates that relevant information in GAN models pre-trained on large-scale source datasets may be transferred to help generation in perceptually different target domains with little data. This transfer is done on both the generator and discriminator [17].

According to testing, all models can accurately represent the correlations between characteristics. The machine learning effectiveness measures show that it is possible to resolve data difficulties for machine learning jobs by substituting synthetic data for actual data [18].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

III. METHODOLOGY

Several data formats can store data generated internally or externally. Different data formats have distinct structures, impacting how quickly they can be processed and how much storage space they need. The choice of data format can affect several factors, including performance speed, compatibility, support for compression, and space requirements.

3.1 Data Description

Based on eight significant aspects, the evaluation, detection and prediction of fake users are being performed. They consist of the user's biography length, the user's following and follower count, the presence or absence of a profile picture, the media count, the length of the username, and the numerical count of the digits present in the username.

1	userBiographyLength
2	userFollowerCount
3	userFollowingCount
4	userHasProfilPic
5	userIsPrivate
6	userMediaCount
7	usernameDigitCount
8	usernameLength

Diagram 1: Classification of each dataset

3.2 Data Visualization

Data Visualization helps us see, understand and interact efficiently with data. This is the representation of data and information in a graphical or pictorial format.

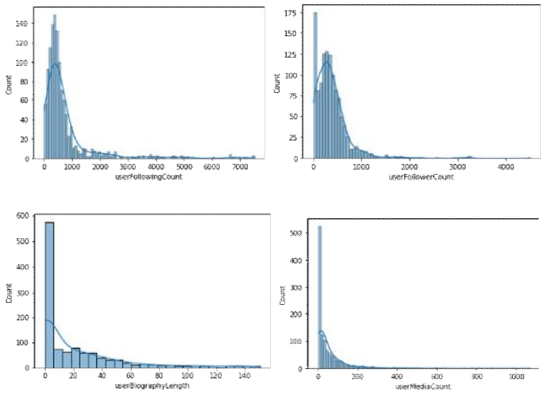

Diagram 2: Histogram

Histograms can help display large amounts of data, data frequency and outliers.

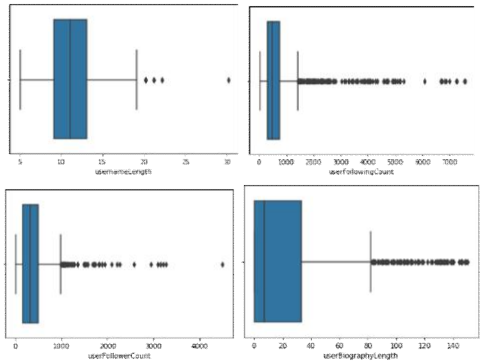

Diagram 3: Boxplot

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

Boxplot helps examine data distribution and can also display mean and standard deviation.

3.3 Data Preprocessing

The first step is converting the format from JSON to CSV. The need to reverse the JSON file to CSV is for better handling of large datasets, has better pre-processing capabilities, requires less storage space, is faster, more secure, and is user-friendly.

There were initially two JSON files, one with fake users and the other with real users. Both the files are converted from JSON to CSV and then merged to create a new data set where the data is trained and generated.

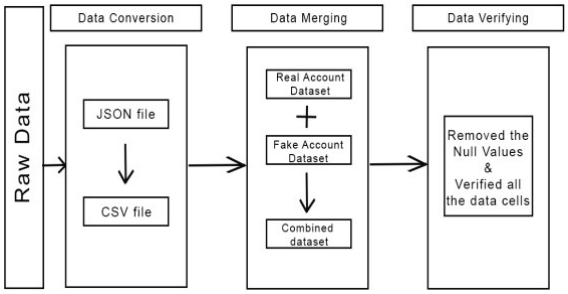

Diagram 4: Processing of Data

After merging, data verification is executed. While merging two extensive datasets, there may be possibilities of the data being duplicated or mislabeled. The data is then verified, and null values are removed. It is made sure no text data is present.

3.4 Data Generation

An estimated 1500 datasets have been collected [1]. Having large data sets helps increase statistical power, improve representation, and produce more accurate results.

For generating new data, we pre-process a neural network model and load data into it. Hence different datasets are constructed by a machine learning algorithm called Generative Adversarial Networks (GANs).

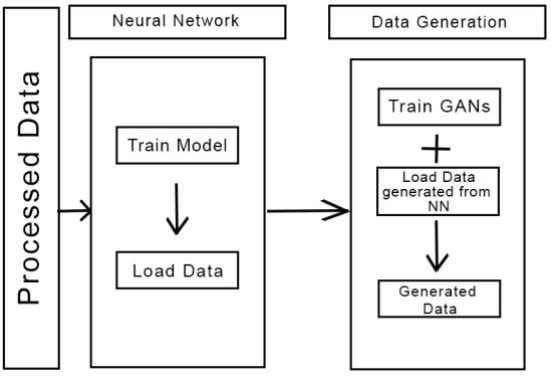

Diagram 5: Data generation using GANs

GANs is a technique for generative modelling that automatically recognizes patterns in data inputs and produces good results based on the actual information.

The GAN has two modules: The generator and the discriminator. The generator trains and generates unique outcomes. The discriminator verifies the fakeness of the outputs.

GANs can only produce numerical data and not text data. Since the dataset used contains only numeric data, it becomes feasible.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

3.5 Data Prediction

A statistical approach called correlation shows how one variable change or moves in connection to another variable. It gives us a general understanding of how closely the two variables are related. Four machine learning algorithms where the data is loaded and then verify the prediction of the dataset.

Logistic Regression - is a supervised algorithm that uses classification to predict the probability of specific classes based on the relationship between the different attributes. The outcome is either 0 or 1; the higher the value, the higher the viability.

Random Forest - is a popular model mainly used for classification and regression problems. The algorithm follows:

- Selecting random samples from the given data.
- Constructing a decision tree for every data.
- Voting by averaging the trees.
- And finally, choose the most voted outcome as the prediction.

XGBoost - is an open-source execution of the popular gradient-boosted trees algorithm. The algorithm follows:

- Make an initial prediction and calculate residuals.
- Construct an XGBoost tree ● Prune the tree ● Compute the output values.
- Make new predictions and calculate the residuals and repeat.

SVM - is a supervised learning Machine Learning algorithm. The objective is to establish the best line or decision boundary to divide several dimensional spaces into classes and quickly classify new data points.

The correlation of each entity in the dataset is observed, and a detailed analysis of the machine-learning algorithms will be used. The datasets are trained and evaluated. Then each algorithm's classification report, confusion matrix, and accuracy are generated and compared. The best and most accurate algorithm is then chosen to predict the true user accounts.

IV. ALGORITHM

Step 1 – Import and converting the two JSON files to CSV.
Step 2 - Merge the two datasets.
Step 3 - Finding the correlation between all the variables to find the relation between each entity.
Step 4 - Visualize and understand the data.
Step 5 - Set the trained data and test data
Step 6 - Generate a larger dataset using GANs and Neural Network.
Step 7 - Using Random Forest, which provides the most significant accuracy.

- Install the required packages
- Load the train data and train the RF model using this data.
- Predict the output on the test data using the trained Random Forest algorithm.

Step 8 - Generate Classification report, Confusion matrix and accuracy to compare.

	Logistic Regression				Random Forest				XGBoost				SVM			
	Precision	Recall	f1-score	Support	Precision	Recall	f1-score	Support	Precision	Recall	f1-score	Support	Precision	Recall	f1-score	Support
True	0.54	1.00	0.70	194	0.93	0.93	0.93	194	0.82	0.98	0.89	194	0.49	1.00	0.66	194
False	1.00	0.18	0.31	204	0.94	0.94	0.94	204	0.98	0.79	0.88	204	0.00	0.00	0.00	204
Accuracy			0.58	398			0.93	398			0.88	398			0.49	398
Macro Avg	0.77	0.59	0.50	398	0.93	0.93	0.93	398	0.90	0.89	0.88	398	0.24	0.50	0.33	398
Weighted Avg	0.77	0.58	0.50	398	0.93	0.93	0.93	398	0.90	0.88	0.88	398	0.24	0.49	0.32	398

Diagram 6: Classification Report Comparisons

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

V. RESULTS

After all the training and fitting the dataset to the model, we have noticed that the training loss we found with the model is very minimum. This indicates that the data has almost perfectly fitted into the deep learning model.

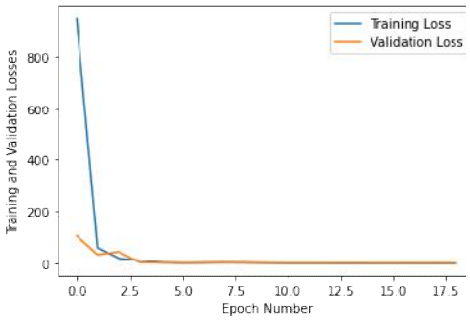

Diagram 7: Model Loss Progression During

5.1 Training/Validation

The performance of the model is very high because the validation loss in this model is also very low, indicating the loss is almost null and making the accuracy the best. With the above visualization, we have also proved that this is an optimized fit model because the training loss and validation loss decrease and meet at the same point in the end.

After running the data with four machine learning algorithms, we found that Random Forest gave us the best results. We found that the accuracy we found with this model to be 93.47%.

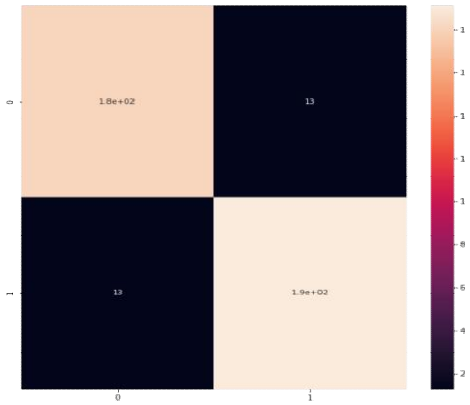

Diagram 8: Confusion matrix of Random Forest

We then plotted the confusion matrix to check the precision, recall and f1-score of the model and be more precise about the model.

Formula,

- Accuracy = (TP+TN)/(TP+TN+FP+FN)
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1-Score = 2*((Precision*Recall)/(Precision*Recall))

After all the calculations, the predictions of all the metrics in the classification report exceeded more the 93% making it one of the best model to prove the trueness of the user using the social media account.

VI. SUMMARY

The evaluation, detection and prediction of fake users are being performed using eight significant aspects: the user's biography length, the user's following and follower count, the presence or absence of a profile picture, the media count, the length of the username, and the numerical count of the digits present in the username.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

The JSON formatted file is converted to CSV, which has better pre-processing capabilities, requires less storage space, is faster, more secure, and is user-friendly. The files are merged, then data verification is executed to ensure no text data is present.

1500 datasets have been collected to increase statistical power, improve representation, and produce more accurate results. GANs is a technique for generative modelling that produces good results based on the actual information. Using this technique, the dataset was increased to approximately 90,000.

We use four machine learning algorithms: Logistic Regression, Random Forest, XGBoost, and SVM. The datasets are trained and evaluated, and the best and most accurate algorithm is chosen to predict the true user accounts.

VII. CONCLUSION

In conclusion, four machine learning techniques are used to research the detection of fraudulent accounts. To our knowledge, this is the first time that GANs has been utilized to produce various datasets.

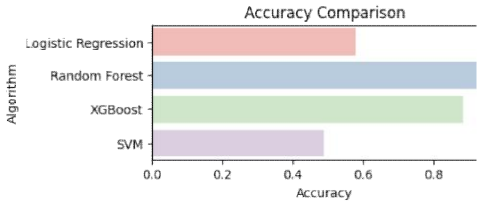

Diagram 9: Accuracy Comparisons

We have contributed the following to this work: gathering the dataset, GANs is used to expand the data, several machine learning methods are proposed, the most accurate algorithm is selected, and time-saving comparisons and results are provided. Research demonstrated that, at 93.47% accuracy, the random forest has the highest accuracy. Moreover, SVM was shown to have the lowest accuracy. We have suggested utilizing the random forest method to identify fake accounts. This can aid in overcoming non-verbal cues and other ineffective strategies.

REFERENCES

[1]. CagatayAkyon, Fatih, and EsatKalfaoglu. "Instagram fake and automated account detection." arXiv e-prints (2019): arXiv-1910.

[2]. Fong, Simon, Yan Zhuang, and Jiaying He. "Not every friend on a social network can be trusted: Classifying imposters using decision trees." In The First International Conference on Future Generation Communication Technologies, pp. 58-63. IEEE, 2012.

[3]. Conti, Mauro, Radha Poovendran, and 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. IEEE, 2012.

[4]. Maniraj, S. P., G. Harie Krishnan, T. Surya, and R. Pranav. "Fake account detection using machine learning and data science." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 9, no. 1 (2019).

[5]. -Cao, Qiang, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. "Aiding the detection of fake accounts in large scale social online services." In Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12), pp. 197-210. 2012.

[6]. Yang, Chao, Robert Harkreader, and Guofei Gu. "Empirical evaluation and new design for fighting evolving twitter spammers." IEEE Transactions on Information Forensics and Security 8, no. 8 (2013): 1280-1293.

[7]. Shama, Sk, K. Siva Nandini, P. Bhavya Anjali, and K. Devi Manaswi. "Fake profile identification in online social networks." Int. J. Recent Technol. Eng 8, no. 4 (2019): 1119011194.

[8]. Fire, Michael, Dima Kagan, AviadElyashar, and Yuval Elovici. "Friend or foe? Fake profile identification in online social networks." Social Network Analysis and Mining 4 (2014): 1-23.

[9]. Jia, Jinyuan, Binghui Wang, and Neil Zhenqiang Gong. "Random walk based fake account detection in online social networks." In 2017 47th annual IEEE/IFIP international conference on dependable systems and networks (DSN), pp. 273-284. IEEE, 2017.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, August 2023

[10]. Tsikerdekis, Michail, and SheraliZeadally. "Multiple account identity deception detection in social media using nonverbal behavior." IEEE Transactions on Information Forensics and Security 9, no. 8 (2014): 1311-1321.

[11]. Lee, Kyumin, James Caverlee, and Steve Webb. "Uncovering social spammers: social honeypots+ machine learning." In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 435-442. 2010.

[12]. Wang, G. Alan, Hsinchun Chen, Jennifer J. Xu, and HomaAtabakhsh. "Automatically detecting criminal identity deception: an adaptive detection algorithm." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 36, no. 5 (2006): 988-999.

[13]. Adikari, Shalinda, and Kaushik Dutta. "Identifying fake profiles in linkedin." arXiv preprint arXiv:2006.01381 (2020).

[14]. Wang, Binghui, Neil Zhenqiang Gong, and Hao Fu. "GANG: Detecting fraudulent users in online social networks via guilt-byassociation on directed graphs." In 2017 IEEE International Conference on Data Mining (ICDM), pp. 465-474. IEEE, 2017.

[15]. Tang, Rui, Luke Lu, Yan Zhuang, and Simon Fong. "Not every friend on a social network can be trusted: an online trust indexing algorithm." In 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 3, pp. 280-285. IEEE, 2012.

[16]. Aziira, A. H., N. A. Setiawan, and I. Soesanti. "Generation of synthetic continuous numerical data using generative adversarial networks." In Journal of Physics: Conference Series, vol. 1577, no. 1, p. 012027. IOP Publishing, 2020.

[17]. Zhao, Miaoyun, Yulai Cong, and Lawrence Carin. "On leveraging pretrained GANs for generation with limited data." In International Conference on Machine Learning, pp. 11340-11351. PMLR, 2020.

[18]. Bourou, Stavroula, Andreas El Saer, Terpsichori-Helen Velivassaki, Artemis Voulkidis, and Theodore Zahariadis. "A review of tabular data synthesis using GANs on an IDS dataset." Information 12, no. 09 (2021): 375.