EXPLORATORY DATA ANALYSIS AND FEATURE SELECTION FOR SOCIAL MEDIA HACKERS PREDICTION PROBLEM

Emmanuel Etuh[1], Francis S. Bakpo[1], George E. Okereke[1] and David Omagu[2]

[1]Department of Computer Science, Faculty of Physical Sciences,
University of Nigeria, Nsukka, Nigeria
[1,2]Department of Mathematics/Statistics, and Computer Science, Kwararafa University,
Wukari, Taraba State, Nigeria

ABSTRACT

In machine learning, the intelligence of a developed model is greatly influenced by the dataset used for the target domain on which the developed model will be deployed. Social media platform has experienced more of hackers' attacks on the platform in recent time. To identify a hacker on the platform, there are two possible ways. The first is to use the activities of the user while the second is to use the supplied details the user registered the account with. To adequately identify a social media user as hacker proactively, there are relevant user details called features that can be used to determine whether a social media user is a hacker or not. In this paper, an exploratory data analysis was carried out to determine the best features that can be used by a predictive model to proactively identify hackers on the social media platform. A web crawler was developed to mine the user dataset on which exploratory data analysis was carried out to select the best features for the dataset which could be used to correctly identify a hacker on a social media platform.

KEYWORDS

User Feature Selection, Prediction, Intrusion Prevention, Hackers detection, Social Media Platform, Machine Learning Algorithm, Training Dataset

1. INTRODUCTION

Social media platforms have become an integral part of average Internet users in the virtual community today. Billions of user devices connected to the Internet operate on one social media platform or the other. According to report in [1], over 500 million Internet of Things (IoT) devices were implemented globally in 2003, 12.5 billion in 2010, and 50 billion in 2020. Online social network platform like Facebook incorporate several functionalities like product and services advertisement and sales that makes it relevant to almost all Internet users. Facebook alone has about 2.89 billion monthly active users as at the second quarter of 2021, Facebook is the biggest social network worldwide[2].

The Covid19 pandemic was instrumental to the geometric shift to virtual socialization, over 56% of the active social media user spend about 43% of their time on social media platform [3]. The technological shift to cloud computing paradigm also has positively influenced the ubiquity of social media. The shift seems to have given hackers an edge to securely carryout their nefarious acts on social media platforms. This has also increased the hacking activity of cybercriminals on the social media platform. According to a survey by Computer Emergency Response Team (CERT), the rate of cyber-attacks have been doubling every year [4]. Online social network is

faced with threatening security challenges [5] because in the digital world, several devices and platforms interact and operate as a family. There are numerous risks that can be raised between all members of the digital family[6].

Cloud intrusion attacks are set of actions that attempt to violate the integrity, confidentiality or availability of cloud resources [7] on cloud SMNP. The rising drop in processing and Internet accessibility cost is also increasing users' vulnerability to a wide variety of cyber threats and attacks. Intrusion Detection System (IDS) is meant to detect misuse, that is, an unauthorized use of the computer systems or its resources by internal and external elements [8]. IDS are effective security technology, which can detect, prevent and possibly react to the attack [9], Shanmugam and Idris in [10] opined that artificial Intelligence plays a driving role in security services like intrusion detection. Several attacks have been launched by these hackers on social media platform[11].

To develop an intelligent system that can efficiently detect hackers on the social media platforms using machine learning approach, the dataset plays a major role. These dataset can be either user features or user activities on the platform. Accuracy of any developed model is dependent on the validity of the dataset used to train the model. Generating dataset from the social media domain mostly employ the use of web crawler that mine information of interest. This is followed by the preprocessing of the dataset for training and testing of the model.

2. LITERATURE REVIEW

To secure the social media platform, many authors have done commendable work on intrusion detection system using it to detect several anomalies on the social media platform. Some of the works on Twitter Bot Detection are that of [12], [13], [14], [15], [16], [17], [18]; network intrusion detection systems have been proposed by [19], [8], [10]; Data warehousing and data mining techniques for intrusion detection systems have been proposed by [4], [9], [20], [21]; Social Media Cyberbullying Detection [22], [23]; Fake account detection [17]. All of these models were developed using dataset to train and test them. Most of the proposals used existing generic dataset in the development of the models which might affect the accuracy of the real-time implementation.

The "juicy prospect" of social media network platform has made hackers to constantly device techniques to intrude and usurp users. They have two fold targets which are the social media users and the SMNP which they break into and control for their selfish gain [24]. On the users end, the hackers' activities make them susceptible to threats which include identity theft, evil twin, password resetting, sim cloning, brute force, fake links, phishing, information leakage, celebrity spoofing, fake account, impersonation, [25] and [26]. They also use code injection through malicious SQL script to disrupt the network.

There are two possible approaches to obtaining machine learning dataset for a model design in the social media domain. The first approach is to use existing dataset while the second approach is to using data mining approach. There are several machine learning datasets available in different repositories which can be used for model development. Some of the popular dataset used in twitter domain and their users include: [13] used detecting-twitter-bot-data from Kaggle; [19] used dataset generated by Cresci in 2017; [12] used TwiBot-20 dataset for their model. In using existing dataset, [12] stated that "low user diversity, data scarcity, and limited user information are the main problems encountered" in most of them.

On the other hand, to obtain the state-of-the-art dataset, data mining approach is often used. Data mining is a data gathering process where the researcher goes to the domain of research to collect data needed for the research. In artificial intelligence domain, data can be mined by developing a web crawler that can be programmed to collect dataset of interest from the domain. Researchers like [27] and [28] used data mining approach. This research work will adopt the approach of data mining to generate the dataset for the model development.

3. MATERIALS AND METHOD

In this research, data mining approach was used for data collection. A Twitter crawler was developed to mine dataset from twitter accounts using Twitter REST API. Domain knowledge was used to mine user information from the twitter account to generate the features for the user dataset. Features for extraction as designed in the crawler are user: id, screen name, location, description, url, protected, followers count, friends count, listed count, created at, favourites count, geo-enabled, verified, status count, language, profile use background image, default profile, and default profile image for all users. The design of the model for data mining is presented in Fig 1.

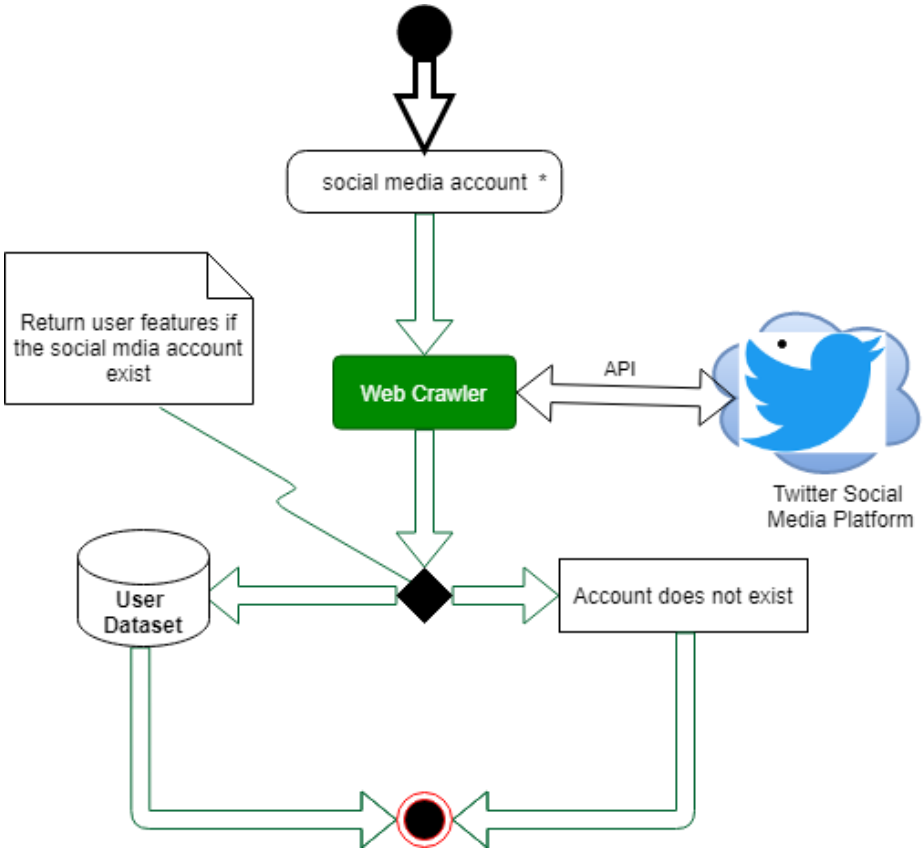

Fig 1. Design of the web crawler for data mining

The generated data set is saved to a .csv file for data wrangling and subsequent analysis.

3.1. Data Pre-processing

About 36,590 dataset was mined for the training and evaluation of the model. Each dataset contain 20 columns that represent features for each twitter account mined. Some of the features would have to be transformed to machine understandable format to make it suitable for the model training and evaluation, data wrangling will be performed on the collected data.

In pre-processing the data, some features that are Boolean can be converted to integer to enable the machine learning algorithm to process them efficiently. Features like protected, geo-enabled, verified, profile_use_background_image, default_profile, and default_profile_image that are represented by Boolean values of true or false will be converted to data type suitable for processing by the model. These Boolean data are converted to its integer equivalence where 0 represent false for each feature while 1 representing true will mean that the feature is present. For example, for the feature protected for a typical user, 1 will mean the user is protected while 0 will mean the user is not protected. Similarly, users that have default_profile on their account will have it represented by integer 1 while those that do not have default_profile will have integer 0 representation. Python built-in function astype(int) is used for the Boolean to integer conversion of these features. The sample of the dataset after converting from boolean to integer is shown in Fig 2.

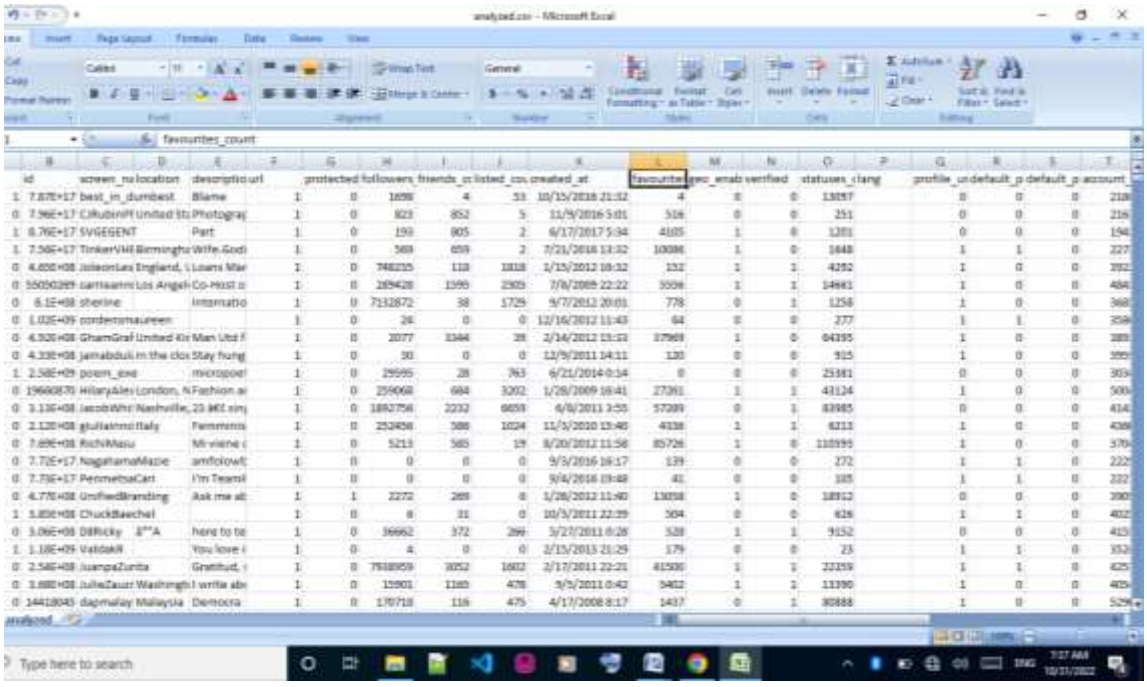

Fig 2. Boolean to integer pre-processing

3.1.1. Visualization of Feature Distribution

Using some inbuilt functions, more features can be deduced form the original 20 features to describe the user in a more explicit manner. First a new feature to see how many days the account has been in use will be created using datetime.datetime.now() function in python programming language. By simple mathematical operations, other features like ‘account_age', 'avg_daily_followers', 'avg_daily_friends', 'avg_daily_favorites' are generated to enhance explicit analysis of a typical user. The full features to be used for account-level verification of user account are: is_hacker', 'id', 'screen_name', 'location', 'description', 'url', 'protected',

'followers_count', 'friends_count', 'listed_count', 'created_at', 'favourites_count', 'geo_enabled', 'verified', 'statuses_count', 'lang', 'profile_use_background_image', 'default_profile', 'default_profile_image', 'account_age', 'average_tweets_per_day', 'hour_created', 'avg_daily_followers', 'avg_daily_friends', 'avg_daily_favorites', 'friends_log', 'followers_log', 'favs_log', 'avg_daily_tweets_log', 'network', 'tweet_to_followers', 'follower_acq_rate', 'friends_acq_rate', 'favs_rate'

3.1.2. Statistical Distribution of Features

The collected features of the dataset are plotted to see the statistical distribution of each feature. If the selected features are skewed, it will drastically affect the accuracy of the model prediction. The following features are plotted on the chart to see their distribution. They are: 'followers_count', 'friends_count', 'listed_count', 'favourites_count', 'statuses_count', 'lang', 'account_age', 'avg_daily_followers', 'avg_daily_friends', 'avg_daily_favorites', 'friends_log', 'followers_log', 'favs_log'.

The distribution of the features are shown in Fig 3 to Fig 14. Most of the distribution graph reveals skewed distribution which signals tendency of over fitting in the model development. For instance, followers' counts of the account users in the dataset plotted in Fig 3 shows a skewed distribution of this feature where the users in the dataset collected has less than 0.1.

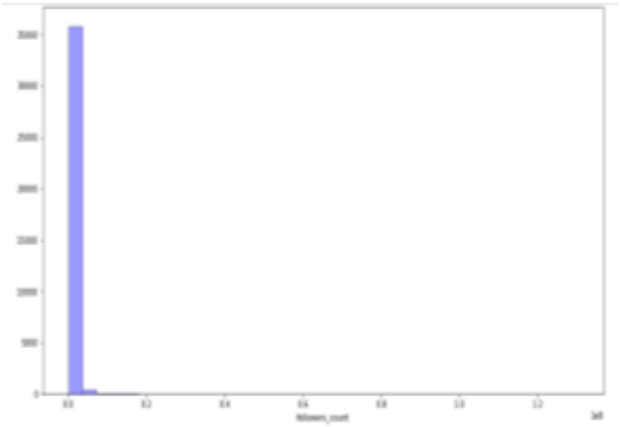

Fig 3. Distribution of followers count

The distribution of the friends' counts for the account users in the dataset plotted in Fig 4 also shows that the distribution is skewed.

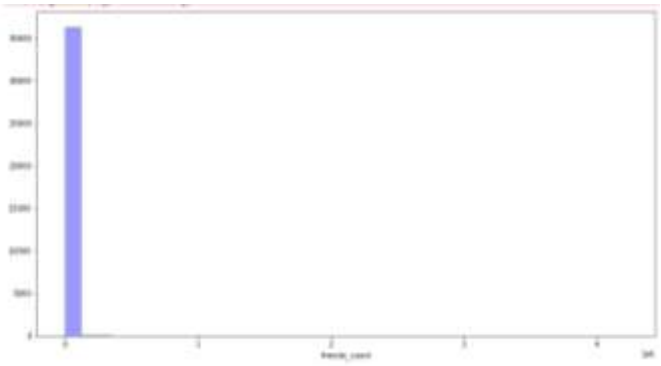

Fig 4. Distribution of friends count

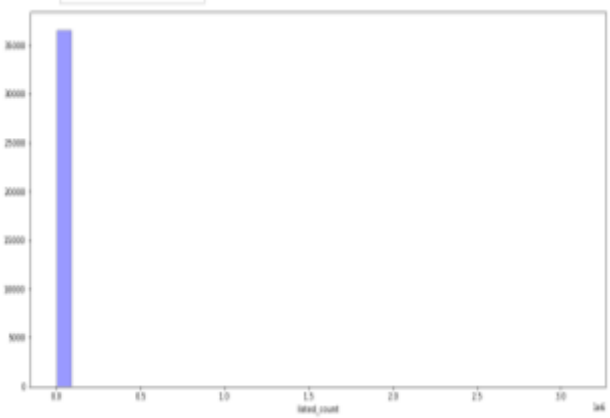

Fig 5. Distribution of listed count

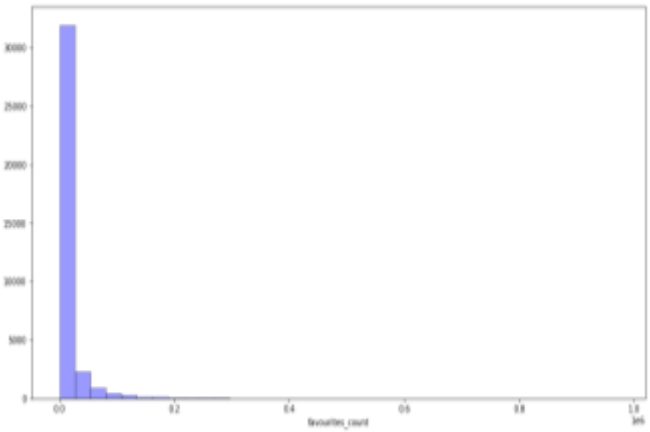

Fig 6. Distribution of favourites count

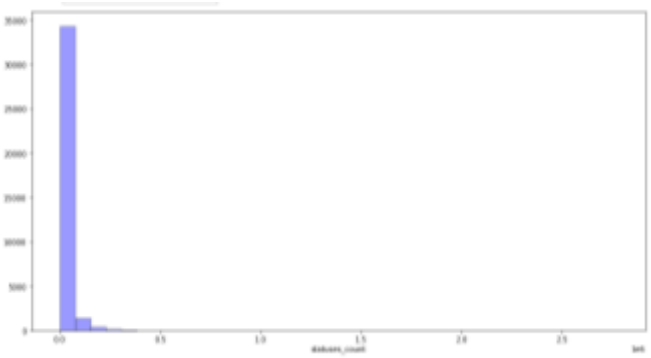

Fig 7. Distribution of status count

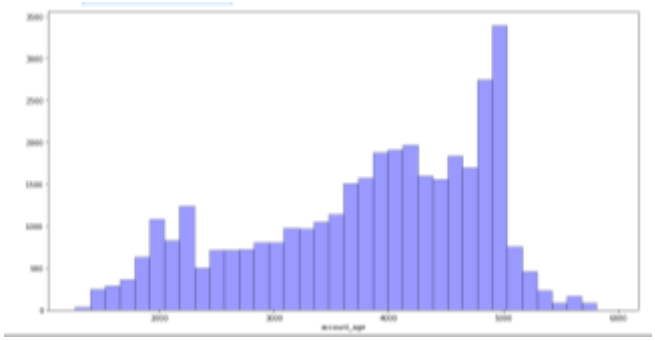

Fig 8. Account age

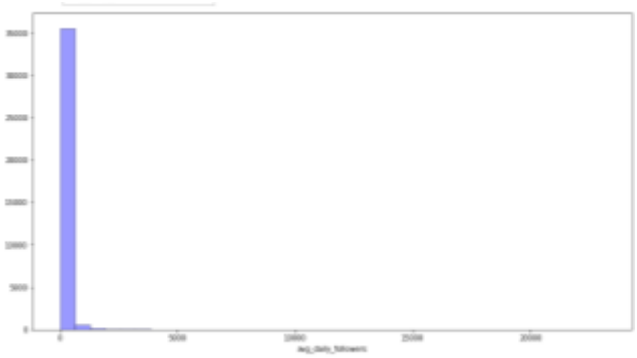

Fig 9. Average daily followers

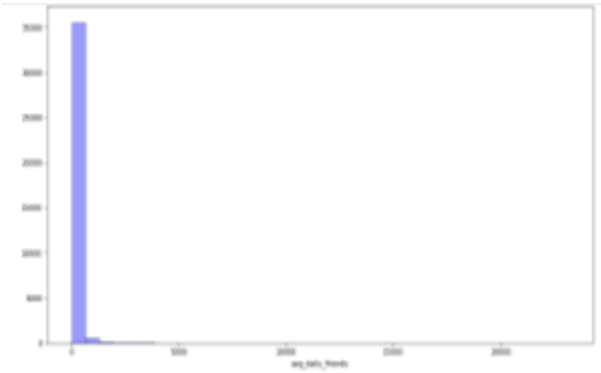

Fig 10. Average daily friends

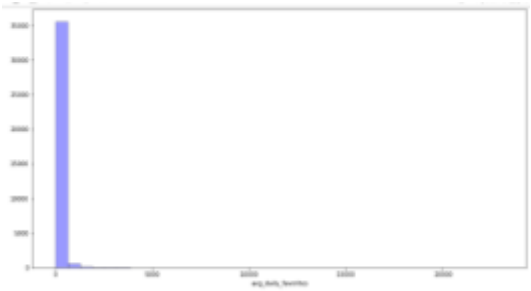

Fig 11. average daily favourites

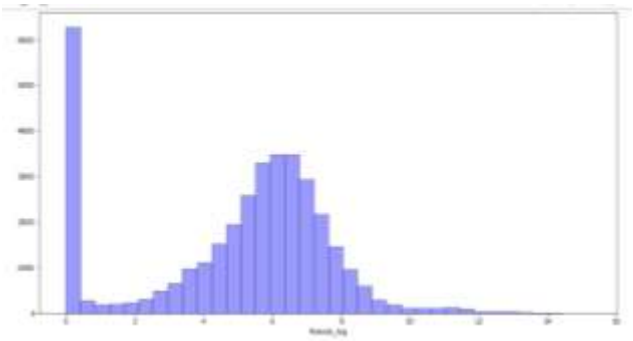

Fig 12. Friends log

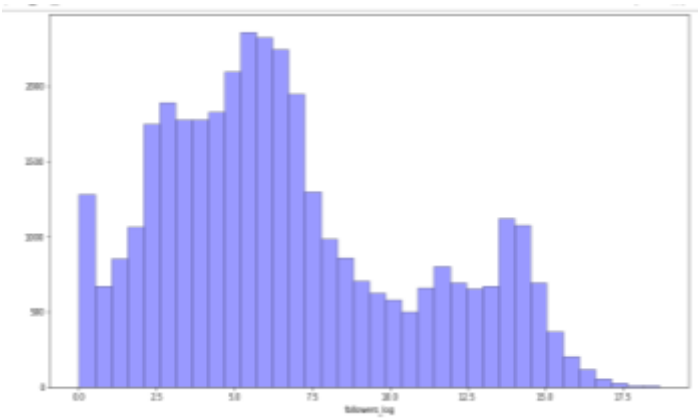

Fig 13. Followers' log

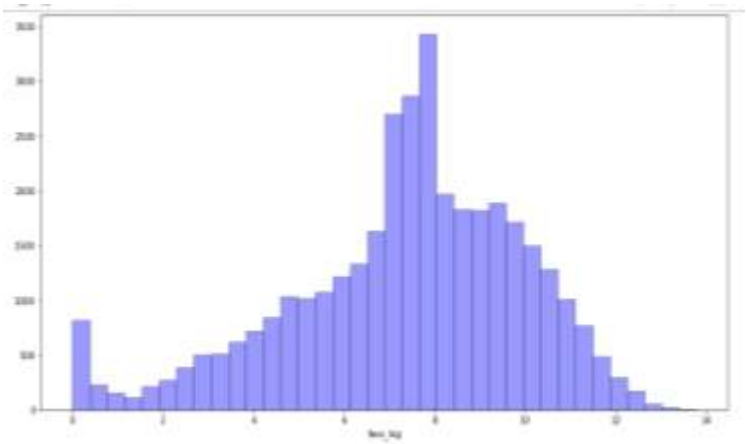

Fig 14. Favourites log

3.1.3. Transformation of Skewed Features

Looking at the distribution of the values for the plots in section 2.1.2, we can see that the values of most of the features are skewed. The transformation of parameters will be carried out to try to correct the skewed features. In the transformation, average_tweets_per_day is gotten by dividing statuses_count' by account_age; hour_created is gotten by multiplying created_at with dt.hour. Other interesting features generated are: 'avg_daily_followers which is gotten by dividing

'followers_count' by 'account_age'; 'avg_daily_friends' which is gotten by dividing the 'followers_count' by 'account_age'; and 'avg_daily_favorites which is gotten by dividing 'followers_count' by 'account_age'.

Logarithmic transformations for highly skewed data are performed for 'friends_log', 'followers_log', 'favs_log', 'avg_daily_tweets_log', and 'average_tweets_per_day'.

Other possible interaction features needed for user identification will be the 'network', 'tweet_to_followers'. Daily acquisition metrics like 'follower_acq_rate, 'friends_acq_rate', and 'favs_rate' will also require a logarithmic transformation of the features.Fig 15 shows the sample of the user dataset after the transformation on the dataset.

orites	friends_log	followers_log	favs_log	avg_daily_tweets_log	network	tweet_to_followers	follower_acq_rate	friends_acq_rate	favs_rate
1.0	1.609	7.438	1.609	1.941	11.968	70.512	0.573	0.002	0.002
0.0	6.749	6.714	6.248	0.110	45.313	37.126	0.321	0.331	0.331
0.0	6.692	5.268	8.320	0.480	35.253	37.358	0.094	0.345	0.345
0.0	6.492	6.346	9.219	0.543	41.198	47.008	0.223	0.254	0.254
190.0	4.779	13.525	5.030	0.738	64.636	113.137	5.254	0.030	0.030

Fig 15. Head of transformed dataset

3.1.4. Correlation of the Features

The confusion matrix is plotted using the feature set to see the correlation of features for hackers only.Fig 16 shows the plot for hacker users only.

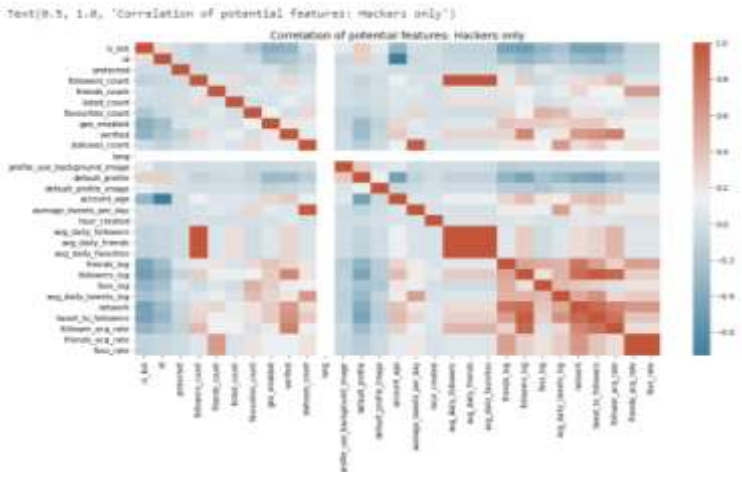

Fig 16. Correlation of potential features for hackers

The confusion matrix is plotted using the feature set of to see the correlation of features for humans only. Fig 17 shows the plot for human users only.

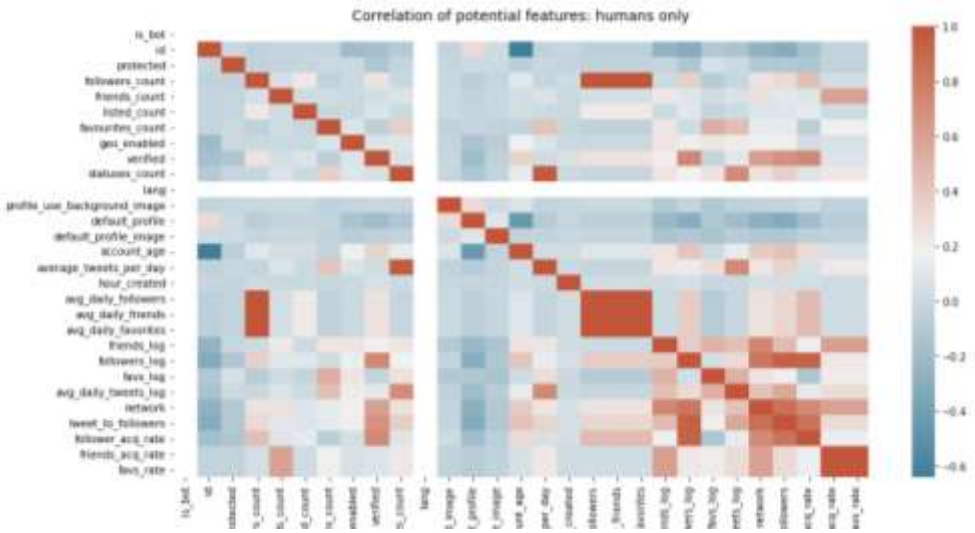

Fig 17. Correlation of potential features: humans only

Fig 18 shows another important feature that could indicate if an account belongs to a human or hacker. This feature is “verified”, that is, if the user account has been verified by the service provider of the social media platform. the graph shows that most of the generated dataset were unverified.

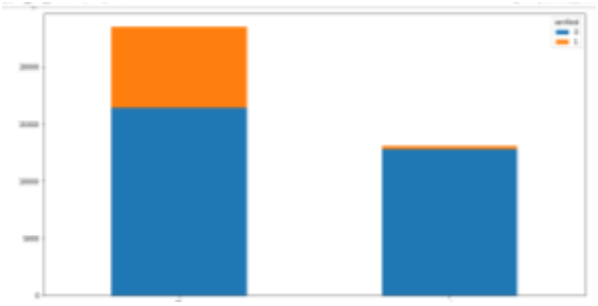

Fig 18. verified versus non-verified accounts

3.1.5. Distribution graph for Hacker and Human users

In this section, the graph to show the statistical distribution of hackers against human are presented.

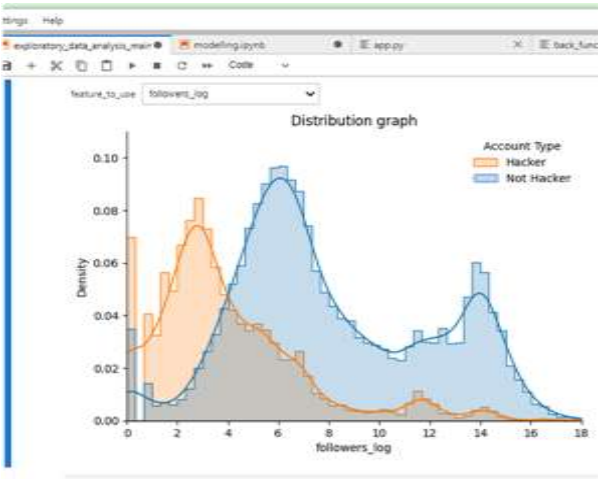

Fig 19. Followers log distribution graph

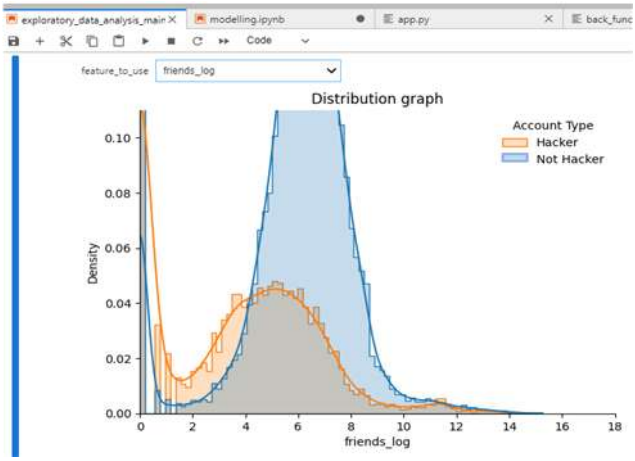

Fig 20. Friends log distribution graph

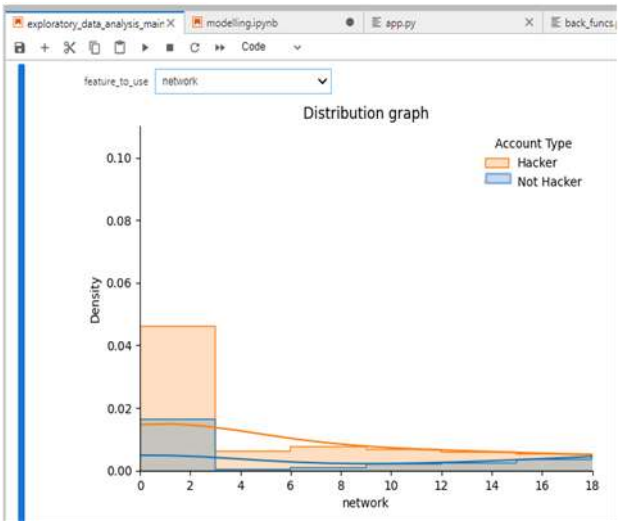

Fig 21. Network distribution graph

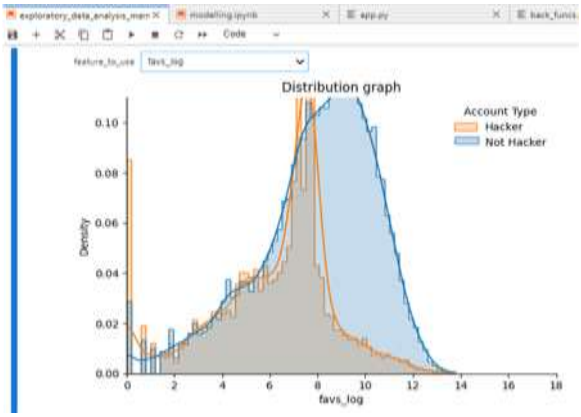

Fig 22. Favourites log distribution graph

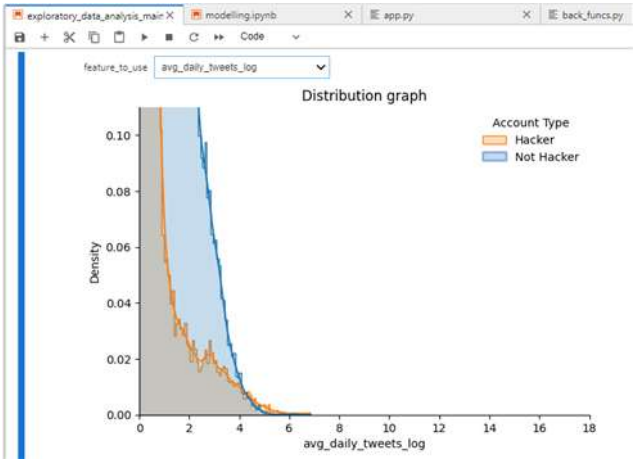

Fig 23. Average daily tweets distribution graph

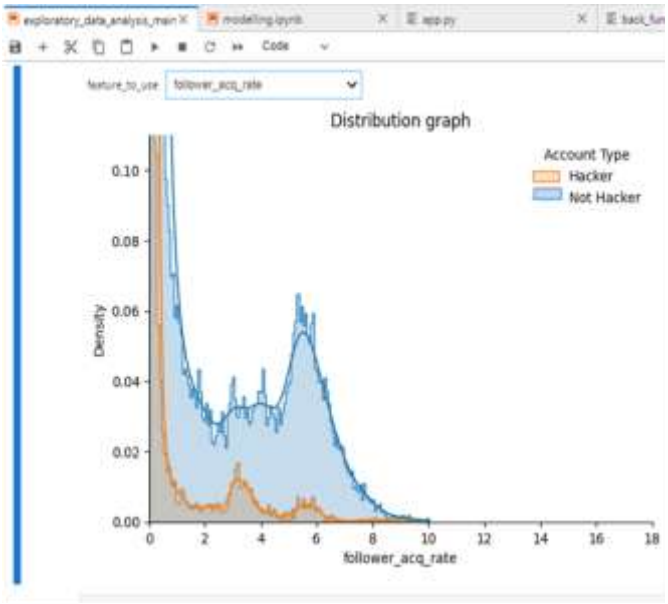

Fig 24. Follower acquire rate distribution graph

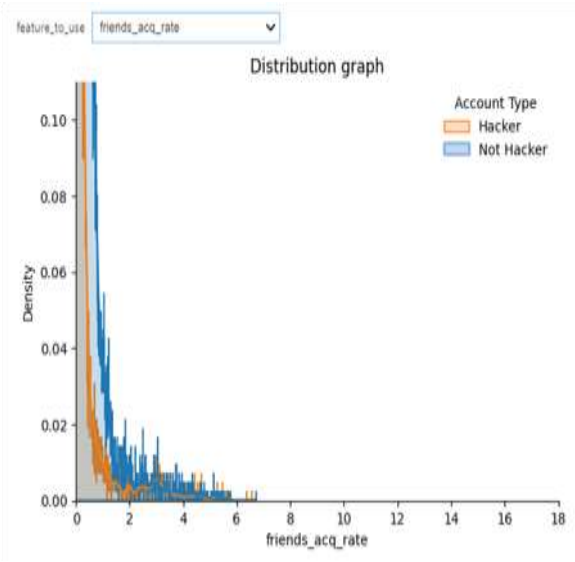

Fig 25. Friends acquire rate distribution graph

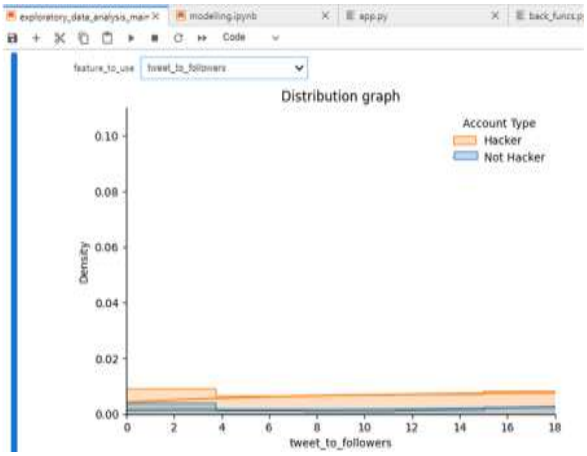

Fig 26. Tweet to followers distribution graph

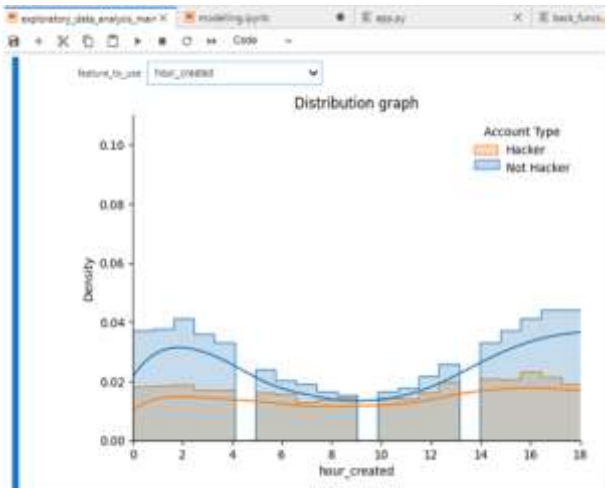

Fig 27. Hour created distribution graph

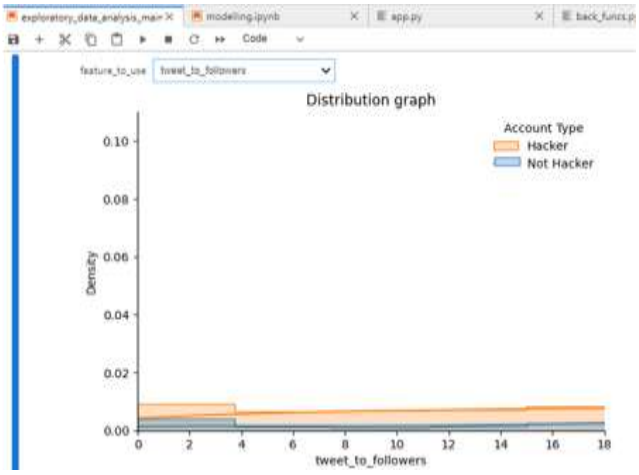

Fig 28. Hour created distribution graph

3.1.6. Dataset Features

The selected features for prediction of a social media user are the 'screen_name', 'created_at', 'hour_created', 'verified', 'geo_enabled', 'lang', 'default_profile', 'default_profile_image', 'favourites_count', 'followers_count', 'friends_count', 'statuses_count', 'average_tweets_per_day', '**account**_age', 'avg_daily_followers', 'avg_daily_friends', 'avg_daily_favorites', 'friends_log', 'followers_log', 'favs_log', 'avg_daily_tweets_log', 'network', 'tweet_to_followers', 'follower_acq_rate', 'friends_acq_rate', and the 'favs_rate'. The sample of the datatset features is shown in Fig 29.

```
d.read_csv("../Exploratory Data Analysis/analyzed.csv", sep=',', lineterminator='\n')
d()
```

	id	screen_name	location	description	url	protected	followers_count	friends_count	listed_count	...	avg_daily_favorites	frien
	787405734442958848	best_in_dumbest	NaN	Blame @xalax, inspired by @MakingInvisible_us..	1	0	1698	4	53	...	1.0	
	796216118331310080	CJRubinPhoto	United States	Photographing the American West since 1980. I...	1	0	823	852	5	...	0.0	
	875948740503859200	SVGEGENT	NaN	Part savage all gentleman □ \n...	1	0	193	805	2	...	0.0	
	756119643622735872	TinkerVHELPIS	Birmingham, AL	Wife.Godmother.Friend.Feline Fanatic! Assistan..	1	0	569	659	2	...	0.0	
	464781334	JoleonLescott	England, United Kingdom	Loans Manager @mancity, @pumafootball Ambassad..	1	0	748235	118	1618	...	191.0	

14 columns

Fig 29. Dataset features

3.1.7. Final Dataset

The dataset analyzed and saved for the model contains 36590 distinct users with 15 features that can be used to predict the type of user on the social media platform. Fig 30 shows the sample of the final dataset used for the model development.

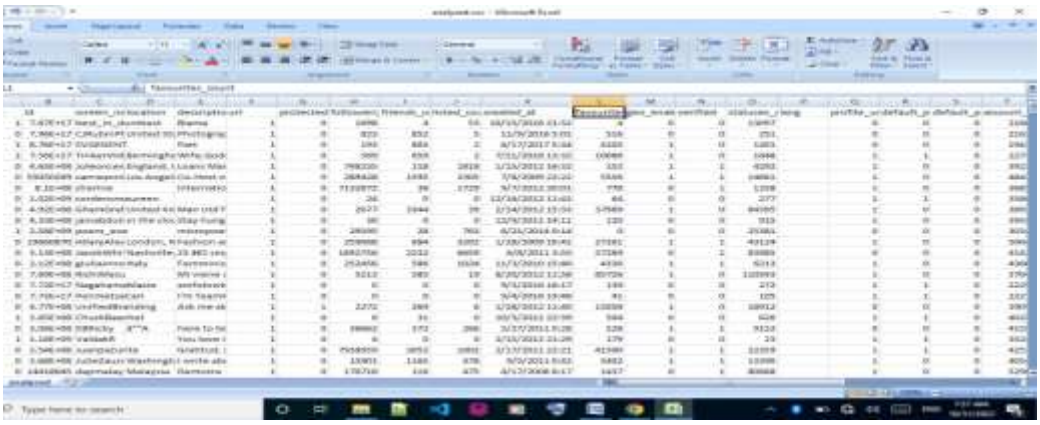

Fig 30. Sample of final dataset

4. DISCUSSION

To accurately predict a social media user, from the exploratory analysis, not all features are relevant. The dataset analyzed and saved for model development contains a dataset of 36590 distinct users with 15 features that can be used to predict the type of user on the social media platform. The final dataset is available with the user and can be used for development of any predictive model or classification problem that has to do with hackers on social media platform.
Two machine learning algorithms were used to validate the dataset. They are Random Forest and XGBoost. To view the feature importance to the two algorithms, charts are plotted to see how each model uses the features; this will help in the understanding of the relevant features necessary for identification of hackers on social media platform.

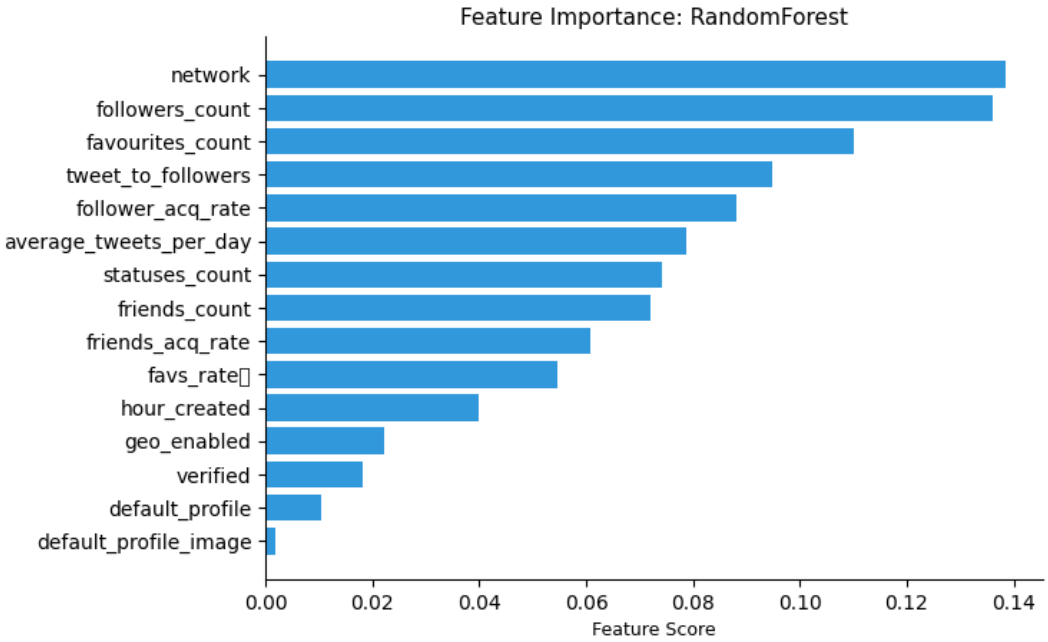

Fig 31. Feature importance for Random Forest

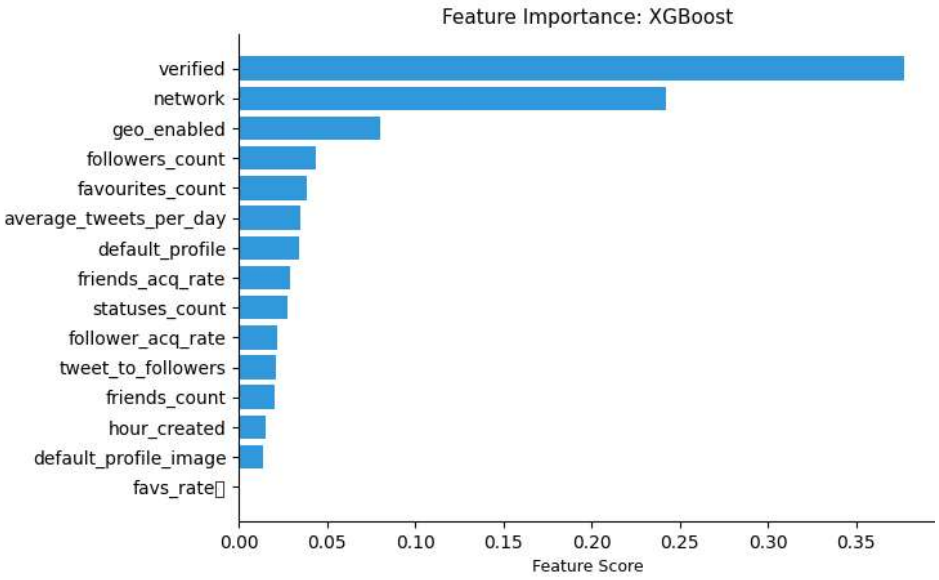

Fig 32. Feature importance for XGBoost

The two figures Fig 31 and Fig 32 show how each model used the features in social media user identification. Network is important for both, but the rest is pretty scattered. Favourites count seems pretty important in both, so does average tweets per day. Random Forest doesn't have verified as more important. Also worthy of note is that the feature scores in Random Forest are much more than that of XGBoost where network and verified are very high and the rest are much smaller.

5. CONCLUSION AND FUTURE WORK

Human physical features or activities can be used to identify a friend in the actual world. To develop an intelligent system that can identify hackers in the virtual world, particularly on the social media platform, user features have been used. To adequately identify a social media user as hacker, there are relevant user details called features that have been used to determine whether a social media user is a hacker or not. In this paper, an exploratory data analysis was carried out to determine the best features that can be used by a predictive model to identify hackers on the social media platform.

A web crawler was developed to mine the user dataset on which exploratory data analysis was carried out to select the best features for the dataset. The data set can be used for any predictive or classification model to separate human users from hackers on social media platform.In the future, we hope to combine the user features with the user activities on the platform to test if it will improve the accuracy of prediction. The dataset is available on request by contacting the corresponding author.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the authors of the works cited in this paper.

REFERENCES

[1] O. Logvinov, "Standard for an Architectural Framework for the Internet of Things (IoT)," 2021. .

[2] P. Jucevi and G. Valinevičienė, "A Conceptual Model of Social Networking in Higher Education," *Electron. Electr. Eng.*, vol. 6, no. (102, 2010.

[3] D. Chaffey, "Global social media statistics research summary 2023," 30-Jan-2023. [Online]. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. [Accessed: 02-Feb-2023].

[4] A. Singhal and S. Jajodia, "Data warehousing and data mining techniques for intrusion detection systems," *Distrib Parallel Databases*, vol. 20, pp. 149–166, 2006.

[5] K. Musial and P. Kazienko, "Social networks on the Internet," *World Wide Web*, pp. 31–72, 2012.

[6] S. K. Assayed, "A C YBERSECURITY AND D IGITAL R ISK A SSESSMENT : A F AMILY C ASE S TUDY," *Comput. Sci. Eng. An Int. J.*, vol. 13, no. 2, pp. 35–41, 2023.

[7] Z. Umar and E. Etuh, "A Framework for Digital Forensic in Joint Heterogeneous Cloud Computing Environment," *J. Futur. Internet*, vol. 3, no. 1, pp. 1–11, 2019.

[8] G. N. Prabhu, K. Jain, N. Lawande, Y. Zutshi, R. Singh, and J. Chinchole, "Network Intrusion Detection System," *Int. J. Eng. Res. Appl.*, vol. 4, no. 4, pp. 69–72, 2014.

[9] H. Vora, J. Kataria, D. Shah, and V. Pinjarkar, "Intrusion Detection System for College ERP System," *J. Res.*, vol. 03, no. 02, pp. 69–72, 2017.

[10] B. Shanmugam and N. B. Idris, "Artificial Intelligence Techniques Applied To Intrusion Detection," in *Proceedings of the Postgraduate Annual Research Seminar*, 2005, pp. 285–287.

[11] E. Etuh, F. S. Bakpo, and A. H. Eneh, "Social Media Network Attacks and Their Preventive Mechanisms: A Review," *Comput. Sci. Inf. Technol. (CS IT)*, vol. 11, no. 4, pp. 59–72, 2021.

[12] L. Rovito, L. Bonin, L. Manzoni, and A. De Lorenzo, "An Evolutionary Computation Approach for Twitter Bot Detection," *Appl. Sci.*, vol. 5915, no. 12, pp. 1–25, 2022.

[13] V. Calleja-solanas *et al.*, "Twitter bot detection using supervised machine learning," *J. Phys. Conf. Ser.*, vol. 1950, no. 2021, pp. 1–11, 2021.

[14] D. Kosmajac and V. Keselj, "Twitter Bot Detection using Diversity Measures," in *3rd International Conference on Natural Language and Speech Processing*, 2019, pp. 1–8.

[15] F. Wei and U. T. Nguyen, "Twitter Bot Detection Using Bidirectional Long Short-term Memory Neural Networks and Word Embeddings," in *First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications*, 2019, pp. 101–109.

[16] P. G. Efthimion, S. Payne, and N. Proferes, "Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots," *SMU Data Sci. Rev.*, vol. 1, no. 2, 2018.

[17] A. El Azab, A. M. Idrees, M. A. Mahmoud, and H. Hefny, "Fake Account Detection in Twitter Based on Minimum Weighted Feature set," *Int. J. Comput. Inf. Eng.*, vol. 10, no. 1, pp. 13–18, 2016.

[18] E. Etuh, G. E. Okereke, D. U. Ebem, and F. S. Bakpo, "A Conceptual Framework of a Detective Model for Social Bot Classification," *Int. J. Ambient Syst. Appl.*, vol. 10, no. 4, pp. 9–17, 2022.

[19] S. Kudugunta and E. Ferrara, "Deep Neural Networks for Bot Detection," *Inf. Sci. (Ny).*, vol. 467, pp. 312–322, 2018.

[20] A. Arora and A. Gosain, "Intrusion Detection System for Data Warehouse with Second Level Authentication," *Int. J. Inf. Technol.*, vol. 13, pp. 877–887, 2021.

[21] R. J. Santos, J. Bernardino, and M. Vieira, "DBMS Application Layer Intrusion Detection for Data Warehouses," in *Building sus- tainable information systems.*, 2013.

[22] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social Media Cyberbullying Detection using Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 703–707, 2019.

[23] B. Shah, N. Sharma, and S. Bandgar, "Cybercrime Prevention on Social Media," *Int. J. Eng. Res. Technol.*, vol. 10, no. 03, pp. 509–513, 2021.

[24] C. Noonan and A. Piatt, *Global Social Media Directory*, no. October. USA: U.S. Department of Energy, 2014.

[25] H. Wilcox and M. Bhattacharya, "A Human Dimension of Hacking : Social Engineering through Social Media," in *IOP Conference Series: Materials Science and Engineering*, 2020.

[26] J. Patterson, "Hacking: Beginner to Expert Guide to Computer Hacking, Basic Security, and Penetration Testing (Computer Science Series)," 2021. .

[27] M. Kantepe and M. C. Ganiz, "Preprocessing Framework for Twitter Bot Detection," in *International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 630–634.

[28] S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, "Profile Characteristics of Fake Twitter Accounts," *Big Data Soc.*, no. December, 2016.

AUTHORS

Emmanuel Etuh is a lecturer in the department of Mathematics/Statistics, and Computer Science at Kwararafa University, Wukari, Nigeria and currently pursuing a PhD degree in Computer Science at the University of Nigeria, Nsukka. He obtained his first degree certificate in Computer Science from Kogi State University, Anyigba in 2009 and an MSc degree in Computer Science from Ahmadu Bello University, Zaria in 2014, His research interests include Intelligent Systems (AI), Cyber Security, Cloud Computing, and Software Engineering.

Bakpo Francis S. is a Professor in the Department of Computer Science, University of Nigeria, Nsukka. He joined the Department of Computer Science, University of Nigeria, Nsukka as a Corp member in 1995, retained by the Department in 1996 as lecturer II and progressed to Professor in 2010. He received his Master's degree in Computer Science and Engineering from Kazakh National Technical University, Almaty (formerly, USSR) in 1994 and Doctorate degree in Computer Engineering in 2008 from Enugu State University of Science and Technology, Agbani. Area of Specialization include: computer architecture, computer communications network, Artificial neural network, intelligent software agents and Petri nets theory and applications.

Okereke George Emeka is a senior Lecturer/Researcher, Computer Science Department, University of Nigeria, Director, Computing Centre, Former Head of Department, Computer Science, University of Nigeria. He obtained a Bachelor of Engineering (Hons.) in Computer Science & Engineering from Enugu State University of Science and Technology and a Master of Science degree in Computer science from University of Nigeria. His PhD is in Digital Electronics & Computing from Electronic Engineering Department of University of Nigeria. He joined the services of University of Nigeria in 1998 as a lecturer in Computer Science Department and is currently a Senior Lecturer. Head of Department from 2017 to 2019. His research interest is in Network security, web security, computer forensics, electronic transfers and security, web design and computer architecture/design. George is married with six children.

David Omagu is a lecturer in the department of Mathematics/Statistics, and Computer Science, Kwararafa University, Wukari, Nigeria and currently a Ph.D student in the department of Computer Science at the Federal University, Wukari. He obtained his first degree certificate in Computer Science from University of Calabar, Calabar in 2005 and an M.Sc degree in Computer Science from Obafemi Awolowo University, Ile-Ife in 2014, His research interests is in machine learning and Information System.