ISSN: 2249-7196
IJMRR/April. 2025/ Volume 15/Issue 2s/59-63
Kruttiventi Bhagavan / International Journal of Management Research & Review

FAKE ACCOUNT DETECTION USING MACHINE LEARNING AND DATA SCIENCE

Kruttiventi Bhagavan
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
K.Venkatesh
(Assistant Professor), Master Of Computer Applications, Dnr Collage, Bhimavaram, Andhra Pradesh.

Abstract: ***Nowadays the usage of digital technology have been increasing exponentially. At the same time the rate of malicious users have been increasing. Online social sites like Facebook and Twitter attract millions of people globally. This interest in online networking has opened to various issues including the risk of exposing false data by creating fake accounts resulting in the spread of malicious content. Fake accounts are a popular way to forward spam, commit fraud, and abuse through online social network. These problems need to be tackled in order to give the user a reliable online social network. In this paper, we are using different ML algorithms like Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF) and K-Nearest Neighbours (KNN). Along with these algorithms we have used two different normalization techniques such as Z-Score, and Min-Max, to improve accuracy. We have implemented it to detect fake Twitter accounts and bots. Our approach achieved high accuracy and true positive rate for Random Forest and KNN. Keywords: Data mining, Classification, Logistic Regression, Support Vector Machine, KNearest Neighbours, Random forest, Normalization.***

I. INTRODUCTION

In this contemporary world, people are being dependent on Online Social Networks (OSNs). As many users are attracted and showing more interest to use OSN in their work-life or for personal uses. This gives an opportunity for the spammers to target people by collecting sensitive information by creating fake accounts. Fake accounts are being created in order to hide their identity and to accomplish their targets . Bucket et al. presented a supervised discretization technique known as Entropy Minimization Discretization (EMD) based on attributes, they have used Naïve Bayes algorithm to evaluate the fake accounts in twitter. This technique can even be applied for all OSNs, for this process they proposed their own dataset with 16 attributes. Finally they presented three different evaluation criteria's before discretization and after that is an increase in accuracy, results from 85.5% to 90.41% and they showed that Naïve Bayes works perfectly for discrete values. To identify fake accounts in OSN Naman et al. proposed a different model in a step by step process firstly they gathered the information and cleaned it. Then created some fabricated accounts. Therefore they validated the data and then injected fake accounts by creating new attributes. At this stage supervised ML techniques are applied and finally results are evaluated. In this process they analysed, identified,and abolished fictitious bot accounts. This model helped them to identify fake accounts created by humans from a real one. These activities motivated the researchers to analyse the abnormal activities of Facebook and twitter users by detecting and studying them. Furthermore, in recent years banks and financial providers in the U.S are even analysing Twitter and Facebook accounts before granting the loan . Whereas to create a Robust Fake account detection model Yeh-Chen and Shyhtsun proposed a strategy to analyse user activity by collecting several popular pages which are at an ease to be attacked. They used collector filter accounts to analyse if any malicious activities like spam keywords, extreme promoting a particular company etc. to verify whether they are among the selected group. Then the model is trained. They used three ML algorithms namely Random Forest (RF), C4.5 decision tree algorithm,

Adaptive Boosting by giving a cluster of features as input for testing the model and a rank score is produced as output, which produces the probability to be a fake account. In this model, the RF classifier performed the best according to the correction rate and their model performed well in the real-world without any over-fit problem.In the present scenario, researchers are using various techniques to analyse fake accounts on OSN platforms by using various attributes. Some analysists detected fake accounts in OSN using user profile. Some other analysists detected by using both sentiment analysis and user behaviour. Furthermore, some researchers used ego networks to analyse the clusters in the social networks and even their tweets. Some crawling tools are also used to extract the data which is available publicly . Qiang et al. proposed a new OSN user system known as Sybil Rank which is dependent on ranking the users using social graph. In this model, social relationships are bidirectional which helps to detect fake accounts in large scale OSN's. Mauro et al. [8] proposed a new approach according to an empirical analysis and structure of typical social network interactions and their statistics to detect fake accounts created in OSN's. They analysed from dynamic point using social network graphs within the content of confidentiality threats. Kaur and Singh proposed a wide range of approaches like supervised, semi-supervised and unsupervised methods. Besides this to detect anomalies in data mining and social networking domain they even analysed according to cluster based, proximity based and classification based networks. Yazan et al. presented an Integro which is a robust and scalable defence machine by using distinct classification theory. They mainly analysed on real accounts who accepted the fake requests and the process in the Integro system that takes place from user-level activities with the help of supervised machine learning algorithms. Finally Integro uses probability in order to rank user accounts. Tsikerdekis and Zeadally proposed a method using nonverbal behaviour in order to detect and identify deception in online social media. In this paper they used Wikipedia as an experiment and their method achieved a high detection accuracy than other methods. They even demonstrated how developers and designers had overcome these nonverbal data in analysing the deceit by increasing the reliability in online communications

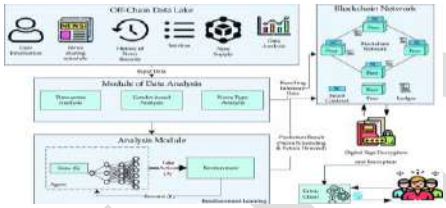
II. LITEARTURE SURVEY

Fake account detection is a critical task in the domain of cybersecurity and social media platforms. The rise of social media has led to an increase in fake accounts, which can lead to several malicious activities, including spamming, phishing, and social engineering attacks. Machine learning and data science techniques have been applied in the past to address this issue, and several research studies have been conducted in this area. Here are a few literature surveys on fake account detection using machine learning and data science.1."A Survey on Detection Techniques for Fake Accounts on Social Networks" by Z. Ji et al. (2018): This survey provides a comprehensive overview of the state-of-the-art fake account detection techniques for social networks. The paper discusses the limitations of traditional detection techniques and presents a range of machine learning-based approaches for detecting fake accounts."A Survey on Fake Account Detection in Online Social Networks" by S. Agrawal et al. (2018): This survey explores the various techniques used for detecting fake accounts in online social networks. The authors discuss several features that can be used to identify fake accounts, such as network features, user behavior, and content analysis. The paper also highlights the limitations of existing approaches and identifies future research directions in this area."Fake Account Detection in Social Media using Machine Learning: A Systematic Literature Review" by H. Lin et al. (2020): This systematic literature review examines the various machine learning techniques used for fake account detection in social media. The authors provide an overview of the most commonly used features and classification algorithms, as well as the datasets used for training and testing. The paper also highlights the challenges of fake account detection and suggests possible solutions.4."Fake User Detection in Social Networks: A Review and Future Directions" by M. Islam et al. (2021): This paper reviews the various techniques used for fake user detection in social networks, including machine learning-based approaches. The authors present a comprehensive survey of the most

commonly used features and classification algorithms, as well as the limitations and future research directions in this area.Overall, these literature surveys highlight the importance of fake account detection and the role of machine learning and data science in addressing this issue. They also provide an overview of the most commonly used techniques, the limitations of existing approaches, and future research directions in this area.

III. PROPOSED METHOD

The existing system uses random forest algorithm identify the fake account.It is efficient when it has the correct inputs and when it has the all the inputs. When some of the inputs are missing it becomes difficult for the algorithm to produce the output. To overcome such difficulties in the proposed system we used a gradient boosting algorithm.Graomponents.dient boosting algorithm is the random forest algorithm which uses decision tree as its main



We also changed the way we find the fake account i.e.,we introduced new method to find the account.The metohods used are spam commenting,engagement rate and artificial activity.These inputs are used to form decision trees that are used in the gradient boosting algorithm.This algorithm gives us an output even if some inputs are missing.This is the major reason for choosing this algorithm .Due to the use of this algorithm we are were able to get highly accurate results.

The proposed work comprises of four steps, (i) data classification, (ii) data preprocessing, (iii) data reduction or transformation, and (iv) Algorithm selection. These steps are described below.

Data Classification

Data classification is characterized as the process for deciding the appropriate type, origin of data and appropriate resources for collecting data. In the data classification step the data is selected from various Twitter accounts. We collected twitter datasets for analysis and to test our model, dataset consists of different attributes such as name, status-count, friend-count and followers-count. We selected these columns as feature attributes status-count, friends-count, followers-count, sex-code, favourites-count, Lang-code.

Data Pre-processing

We used machine learning algorithms in this process to convert and analyse the available raw data into feasible data. It is mainly used for better and accurate results. For instance, some algorithms like Random Forest do not support null values or they need particular format. In such a case data pre-processing is a necessary step. Then we extracted two Comma-Separated Value (CSV) files fake and genuine users, we combined both files by sampling noise in the data then feature labels were added as 0/1 to distinguish fake or real. We selected particular columns as feature attributes status-count, friends-count, followers-count, listed-count, sex-code, favourites-count, Lang-code and then we have removed columns having more null values sex-code, Lang-code, and normalized the data for better accuracy. In this method, we distributed the information for training and testing purpose at 80:20. To represent the values in confusion matrix, the model is trained with x- train and y-train data, and then it is trained with test data for precision and recall values. Graphs are represented as Area under ROC Curve (AUC) and Receiver operating characteristic curve (ROC).

Data transformation

It is a method of converting information from one format or structure into another format. For tasks such as data integration and management, data transformation is a crucial step to improve the accuracy. In our model we used two normalization techniques for data transformation.

Normalization in Data Mining

We are using two data normalization techniques such as Z-Score, Min-Max, it is mainly used when dealing with multiple attributes on different scale and to scale the information into smaller range, it is commonly applied for classification algorithms to

ISSN: 2249-7196
IJMRR/April. 2025/ Volume 15/Issue 2s/59-63
Kruttiventi Bhagavan / International Journal of Management Research & Review

improve the performance rate, so the attributes are normalized to bring on the same scale.

Score: Z-Scores are mainly based on mean value and standardized score these scores are linearly transformed data value with a mean of 0 and the scores have been given a common standard. It helps to understand the rate of a score as per the normal distribution of the data.

Min-Max: In this method linear transformation is performed on original data, according to this minimum values of that feature is transformed as 0 whereas maximum values are transformed into 1 and remaining values have been changed into decimals between 0 and 1.

IV. RESULTS

V. CONCLUSION

Fake account detection is an important problem for many online platforms, as fake accounts can be used for various fraudulent and malicious activities. Machine learning and data science can be used to build an effective fake account detection system, which can help online platforms to detect and remove fake accounts in real-time. The process of building a fake account detection system using machine learning and data science involves several steps, including data collection, preprocessing, feature engineering, machine learning model training, and continuous improvement. The system can be deployed in a production environment, integrated with the platform's registration process, and monitored regularly to ensure that it is performing as expected. The key advantage of using machine learning and data science for fake account detection is the ability to analyze large amounts of data and identify patterns that are indicative of fake accounts. This can help to identify fake accounts that may have otherwise gone undetected, and help online platforms to maintain a safe and secure environment for their users. In conclusion, the use of machine learning and data science for fake account detection is a promising approach that can help to improve the security of online platforms and protect users from fraudulent and malicious activities.

REFERENCES

1) Cao, X., David, MF., Theodore, H.: Detecting Clusters of Fake Accounts in Online Social Networks. In: 8th ACM Workshop on Artificial Intelligence and Security, pp. 91-101 (2015).

2. Buket, E., Ozlem, A., Deniz, K., Cyhun, A.: Twitter Fake Account Detection. In: IEEE 2nd International Conference on Computer Science and Engineering, pp. 388-392 (2017).
3. Naman, S., Tushar, S., Abha, T., Tanupriya, C.: Detection of Fake Profile in Online Social Networks Using Machine Learning. In: IEEE International Conference on Advances in Computing and Communicaton Engineering. pp. 231-234 (2018).
4. Sarah, K., Neamat, E., Hoda, M .O .M.: Detecting Fake Accounts on Social Media. In: IEEE Intenational Conference on Big Data. pp. 3672-3681 (2018).
5. Yeh-Cheng, C., Shyhtsun, F .W.: FakeBuster: A Robust Fake Account Detection by Activiy Analysis. In: IEEE 9th International Symposium on Parallel Architectures, Algorithms and Programming. pp. 108-110 (2018).
6. Myo, MS., Nyein, NM.: Fake Accounts Detection on Twitter using Blacklist. In: IEEE 17th International Conference on Computer and Information and Information Science. pp. 562-566 (2018).
6. Qiang, C., Michael, S., Xiaowei, Y., Tiago P.: Aiding the Detection of Fake Accounts in Large Scale Social Online Services. In: 9th USENIX Conference on Networked Systems Design and Implementation. pp. 1-14 (2012).
7. Mauro, C., Radha, P., Macro, S.: Fakebook : Detecting Fake Profiles in Online Social Networks. In: IEEE International Conference on Advances in Social Networks Analysis and Mining. pp. 1071-1078 (2012).
8. Kaur , R., and Singh, S.: A survey of data mining and social network analysis based anomaly detection techniques. In: Egyptian informatics journal. pp.199–216 (2016).
9. Yazan, B., Dionysios, L., Georgos, S., Jorge, L., Jose, L., Matei, R., Konstatin, B., and Hassan, H.: Integro : Leveraging victim prediction for robust fake account detection in large scale osns ,Computers & Security. pp. 142–168 (2016).
10. Tsikerdekis, M., Zeadally , S.: Multiple Account Identity Deception Detection in Social Media using Non Verbal Behaviour. In: IEEE Transactions on Information Forensics and Security. 9(8), 1311-1321 (2014).
11. How fake news and hoaxes have tried to derail jakarta's election. Internet draft(online).
12. Political advertising spending on facebook between 2014 and 2018 Internet draft.

2) [Online].Available:https://www.statista.com/statistics /891327/politicaladve rtisingspending- facebook by-sponsor-category/.2018