IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Fake Account Detection Using Machine Learning

Balu V[#1], Ananda Raju P[*2]

[#]*Computer Science And Engineering Department, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Deemed University*

Kanchipuram, Tamil Nadu, India

Abstract- In the current generation, online social networks have grown in popularity, which has an impact on how individuals interact with one another on different social media platforms. Several activities, including promotion, communication, agenda formulation, advertisement, and news generation, have started to be done on social media platforms. It has gotten simpler to add new friends and stay in touch with them and their updates. These online social networks have been the subject of research to see how they affect people. Some fraudulent accounts are used to spread false information and further political agendas, for example. Finding a fraudulent account is important. Machine learning-based techniques were used to find fake accounts that might deceive users. The dataset is pre-processed using various python libraries and a comparison model is obtained to get a feasible algorithm suitable for the given dataset. Several MachineLearning algorithms try to identify fraudulent accounts on social media platforms. The detection of fake accounts takes advantage of the classification abilities of the algorithms K-Nearest Neighbors, Naive Bayes, and Support Vector Machines.

Keywords— Classification , Fake accounts ,Supportvectormachine , Random forest .

I. INTRODUCTION

Because it helps to stop the spread of misleading information, fraud, and harmful behaviour, fake account identification is crucial to maintaining the confidence and safety of online platforms. In order to keep one step ahead of fraudsters and safeguard their consumers, online platforms must continue to develop and improve their false account detection algorithms.

To find patterns of behaviour that are compatible with fraudulent activities, fake account detection algorithms use machine learning methods and data analysis. These algorithms may examine

user activity, including login times, posting frequency, language, and connections. They may also examine account information including the profile image, location, and email address.The algorithm can assess this data to determine whether the account is likely to be real or fake. The detection of fake accounts uses the classification capabilities of the algorithms K-Nearest Neighbors, Random Forest, and Support Vector Machines. The mentioned algorithms' accuracy rates for identifying fake accounts are compared, and the algorithm with the highest accuracy rate is noted.

II. LITERATURE SURVEY

In order to provide effective detection for fake Twitter accounts, feature selection, and dimension reduction techniques, Sarah Khaled et al. introduced a new algorithm, SVM-NN. In spite of using fewer features, the proposed algorithm (SVM-NN) can classify 98% of the accounts in our training dataset correctly.

Sreenivas Kumacham et al. proposed a machine learning model to predict the student placements using various Machine Learning algorithms that include J48, Naïve Bayes, Random Forest etc., The model tries to obtain the results from various algorithms and these results are compared to predict the best algorithm for any given dataset.

A hybrid model and skin detection algorithm have been used by M. Smruthi, N. Harini [2] to identify fake accounts on social media. The proposed work's strength is its ability to accurately identify fake accounts. The 400 mixed-five Supervised Machine Learning Algorithm dataset of fake and real accounts was used to produce the results of the proposed work. The accuracy of the Supervised Machine Learning Algorithm is calculated here using 200 fake accounts and 200 real accounts. On the other hand, the algorithm for skin detection is also employed. The image is gathered and calculated for the fake and real accounts based on the percentage of skin exposed.

III. PROBLEM STATEMENT

Misleading accounts can be established with the intent to trick users, disseminate false information, or carry out harmful activities like scams, phishing, and identity theft. To find patterns of behaviour that are consistent with fraudulent activity, the model should examine different aspects of the user account, such as login times, frequency of posting, language used, profile picture, location, and email address. The model should be tested on a holdout dataset to gauge how well it detects fake accounts after being trained on a labelled dataset where each account is classified as genuine or fake.

IV. PROPOSED METHOD

By comparing the accuracy of three machine learning algorithms, the proposed system gathers the dataset that has been preprocessed in order to identify fake Twitter profiles. The algorithm with the highest efficiency is then found for the given dataset. An algorithm can model a problem in a variety of ways depending on how it interacts with its environment or experience during the model preparation process. This interaction allows for the best results to be achieved by selecting the most suitable algorithm for the input data that is provided.

V. ALGORITHM

Support Vector Machine (SVM):

Classification and regression issues are resolved using Support Vector Machine, or SVM, one of the most used supervised learning techniques. Yet it is mostly used for Machine Learning Classification problems. In order to swiftly categorise fresh data points in the future, the SVM algorithm aims to define the best line or decision boundary that can divide n-dimensional space into classes.

LinearSVC: Data that can be separated into two groups using just one straight line are referred to as linearly separable data, and linearly separable data is used in linear SVM. In such cases, a particular kind of classifier called linear SVM is used.

Random Forest:

popular algorithm in machine learning A part of the supervised learning approach is Random Forest. It can be used for ML issues involving both regression and classification. Its foundation is the idea of ensemble learning, a technique for mixing several classifiers to handle difficult problems and improve model performance. As its name suggests, Random Forest is a classifier that averages several decision trees applied to various subsets of the input information to improve the predicted accuracy of the dataset. The random forest uses forecasts from each decision tree and predicts the outcome based on the votes of the majority of projections rather than depending solely on one decision tree.

K-Nearest Neighbour

One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbour. Although the K-NN algorithm is most frequently used for classification problems, it can also be used for regression. Since K-NN is a non-parametric algorithm, it makes no assumptions about the underlying data. It is also known as a lazy learner algorithm because it stores the training dataset rather than learning from it immediately. Instead, it uses the dataset to perform an action when classifying data.The KNN algorithm simply stores the dataset during the training phase and classifies new data into a category that closely resembles the training data.

VI. ARCHITECTURE

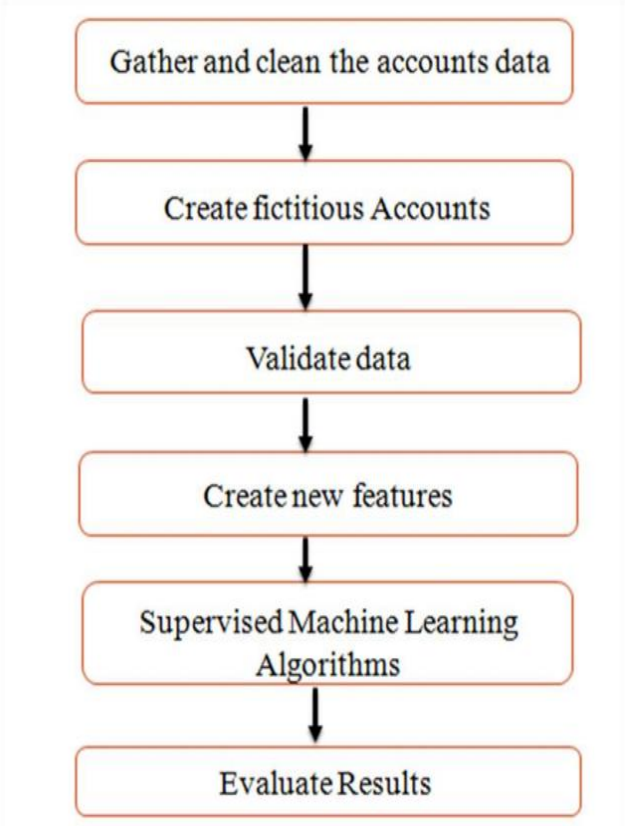

Fig.1 Architecture

i. Data Collection: The first step is to collect data on user accounts from the online platform. This may include information such as login times, frequency of posting, language used, profile picture, location, and email address.

ii. Data Preprocessing: The collected data needs to be cleaned and preprocessed before being used for training the machine learning model. This could include eliminating duplicates, dealing with missing numbers, and preparing the data for analysis.

iii. Feature Engineering: The next step is to extract relevant features from the preprocessed data. This may involve analyzing user behavior and account details to identify patterns of behavior that are consistent with fraudulent activity.

iv. Model Training: Once the features have been extracted, a machine learning model is trained on the labeled dataset. Various machine learning algorithms can be used for this task, such as logistic regression, decision trees, random forests, and deep learning.

v. Model Evaluation: The trained model is evaluated on a holdout dataset to measure its performance in detecting fake accounts. The evaluation metrics may include accuracy, confusion matrix, F1 score.

vi. PROJECT DESCRIPTION

The detection of fake accounts is a significant issue for social media, e-commerce, and online security. False accounts can be created with the intention of doing something bad, like disseminating false information, participating in online bullying, or committing financial fraud. It would be necessary to gather information for the project on various user behaviours and traits that might lead to a fake account. To find a workable solution fit for the provided dataset, the dataset is pre-processed using a variety of Python modules. Several Machine Learning algorithms try to identify phoney accounts on social media platforms. For the purpose of identifying bogus accounts, the classification abilities of the algorithms K-Nearest Neighbors, Random Forest, and Support Vector Machines are applied. The best machine learning algorithm for the job of detecting fake accounts might then be chosen after being tested against a variety of other algorithms. Decision trees, support vector machines, neural networks, and random forests are a few potential algorithms.

VII. RESULT

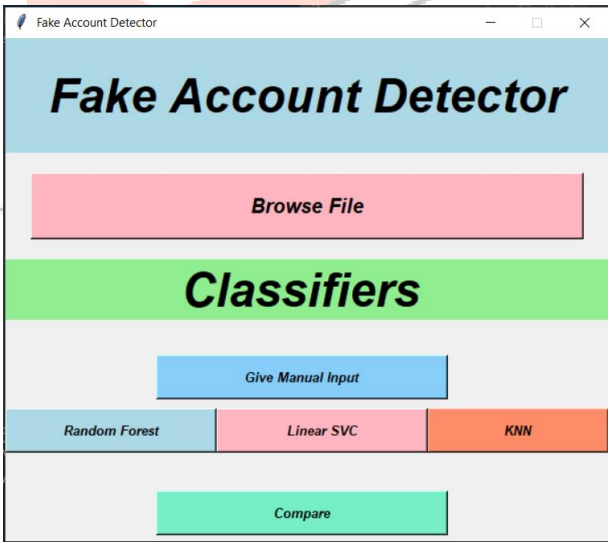

Fig.2.1 The basic interface that shows functionality of project.

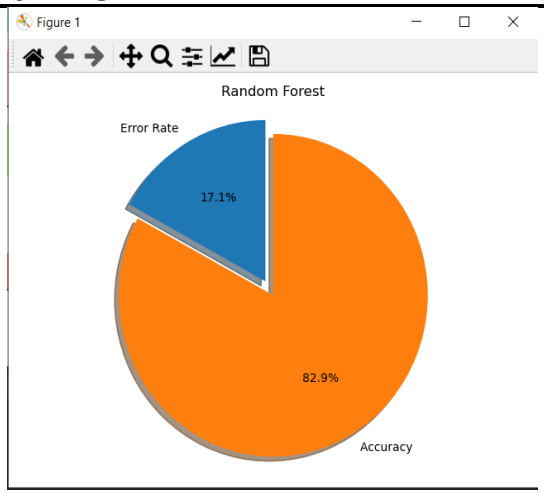

Fig.2.2 Accurracy of Random Forest Algorithm

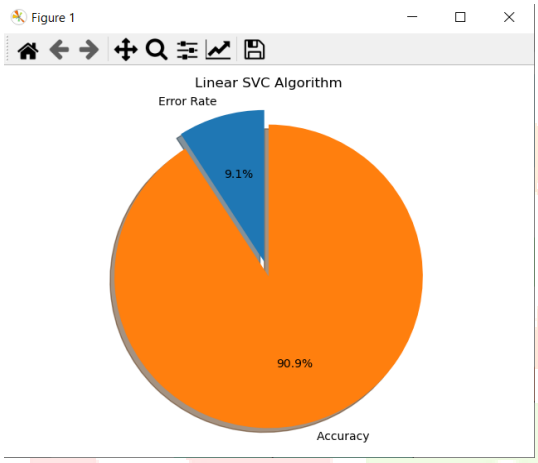

Fig.2.3 Accuracy of support vector machine Algorithm

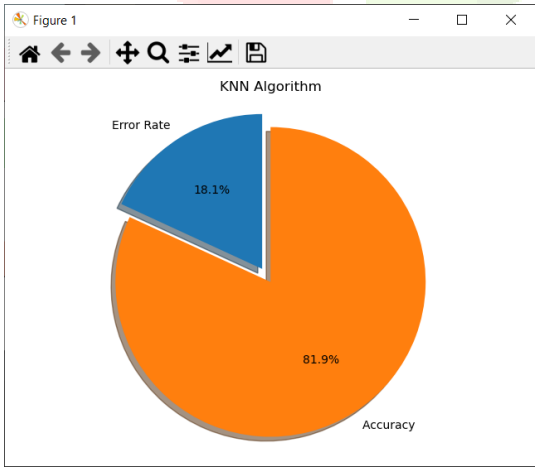

Fig 2.4 Accuracy of K Nearest Neighbour Algorithm

s

Fig 2.5 Algorithm determining the given account details are fake.

VIII. CONCLUSION

This project aims to exploit various dataset aspects that haven't been thoroughly considered in the literature through the use of various machine learning algorithms and to discover a reliable method of identifying automated and fake accounts. In this project, we've demonstrated a machine learning pipeline for spotting fake social network accounts. Our system uses three different classification algorithms to determine whether or not an account in the provided dataset is a fake account, as opposed to making a prediction using just one algorithm. K-Nearest Neighbors, Random Forests, and Support Vector Machine all performed well in our evaluation; however, Support Vector Machine appeared to have the highest prediction accuracy for the given dataset. The Accuracy of detecting fake accounts is found to be higher using SVM Algorithm followed by Random Forest

Algorithm for a given dataset.

IX. REFERENCES

1. Nazir, Atif, Saqib Raza, Chen-Nee Chuah, BurkhardSchipper, and C. A. Davis. “Ghostbusting Facebook:Detecting and Characterizing Phantom Profiles inOnline Social Gaming Applications.” In *WOSN*. 2010.

2. Adikari, Shalinda, and Kaushik Dutta. “Identifying Fake Profiles in LinkedIn.” In *PACIS*, p. 278. 2014.

3. Chu, Zi, Steven Gianvecchio, Haining Wang, andSushil Jajodia. “Who is tweeting on Twitter: human, bot, or cyborg?” In Proceedings of the 26th annual computer security applications conference, pp. 21-
30. ACM, 2010.

4. Stringhini, Gianluca, Gang Wang, Manuel Egele,Christopher Kruegel, Giovanni Vigna, Haitao Zheng, Ben Y. Zhao. “Follow the green: growth and
dynamics in twitter follower markets.” InProceedings of the 2013 conference on Internetmeasurement conference, pp. 163-176. ACM, 2013.

5. Thomas, Kurt, Damon McCoy, Chris Grier, Alek Kolcz,and Vern Paxson. “Traffificking Fraudulent Accounts:The Role of the Underground Market in Twitter Spam and Abuse.” *In* Presented as part of the 22nd

6. Farooqi, Gohar Irfan, Emiliano De Cristofaro, Arik Friedman, Guillaume Jourjon, Mohamed Ali Kaafar,M. Zubair Shafiq, and Fareed Zaffar. “Characterizing Seller-Driven Black-Hat Marketplaces.” arXiv preprint arXiv: 1505.01637 (2015).

7. Viswanath, Bimal, M. Ahmad Bashir, Mark CrovellaSaikat Guha, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. “Towards
detecting anomalous user behavior in online socialnetworks.”

8. S. Khaled, N. El-Tazi and H. M. O. Mokhtar, "Detecting Fake Accounts on Social Media," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 3672- 3681.

9. Rao, K. Sreenivasa, N. Swapna, and P. Praveen Kumar. "Educational data mining for student placement prediction using machine learning algorithms." Int. J. Eng. Technol. Sci 7.1.2 (2018): 43-46.

10. Y. Boshmaf, D. Logothetis, G. Siganos, J. Lería, J. Lorenzo, M. Ripeanu, K. Beznosov, H. Halawa, "Íntegro: Leveraging victim prediction for robust fake account detection in large scale osns", Computers & Security, vol. 61, pp. 142-168, 2016.

11. N. Singh, T. Sharma, A. Thakral and T. Choudhury, "Detection of Fake Profile in Online Social Networks Using Machine Learning," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, 2018, pp. 231-234.

12. D. M. Freeman, "Detecting clusters of fake accounts in online social networks", 8th ACM Workshop on Artificial Intelligence and Security, pp. 91101.