Profile Imposter Detection On Instagram Using XGboost And SVM Algorithm

Shrija Madhu[1], Neeli Gowri Sreelakshmi[2],Gandham Roshitha Madhuri[3], Panapana Shanmukha[4], Dulla Venkata Rajesh[5], Yendluri Venkata Sai Bhanu[6]

[1,2,3,4,5,6]Department of Computer Science and Engineering, Godavari institute of Engineering & Technology,Rajanagaram,Andhra Pradesh, India

[1*]shrija@giet.ac.in, [2] srilakshmi@giet.ac.in,[3]gandhamroshithamadhuri@gmail.com, [4]panapanashanmukha6@gmail.com, [5]20551a0508.rajesh@gmail.com, [6]yendluribhanu0656u@gmail.com

Abstract-In today's world, everyone relies heavily on social media. The vast majority of individuals nowadays regularly use social media as their primary means of communication. Social networking site membership is growing exponentially every day, and many users are chatting with people across the world regardless of time or place. This offers another vector for attack, such as fabricated information. Our study focuses on determining if an Instagram account is real or fake. In order to determine whether newly provided account information is from a legitimate user or an imposter, an algorithm will be trained using historical data on both types of accounts. To identify fake profiles, we employed machine learning methods like XGBoost Algorithm and SVM. The given dataset is pre-processed using multiple different Python tools, and the resulting results are compared to build a realistic method. For the purpose of identifying fake profiles, we compare the results of the classification algorithms XGBoost and SVM Algorithm.

Keywords-XGBoost, SVM, profile imposter detection, Instagram.

1. Introduction

Social networking has become a popular online activity, with millions of people spending billions of minutes on platforms like Facebook, Myspace, Twitter, and Google Buzz. However, protecting user data and fixing security holes in online social network services (OSNs) remain essential goals. . Social media allows individuals from all walks of life to make connections, share ideas, and engage with the data and media available on the network.

Fake accounts, also known as fraudulent or deceptive profiles, represent a multifaceted challenge within the social media platforms. These accounts are primarily created with the intent to deceive, manipulate, or engage in malicious activities, all while assuming the guise of legitimate users. They emerged from a diverse array of places, ranging from automated bots that perform repetitive tasks to impersonation accounts that mimic real individuals, spam and phishing profiles that inundate users with unwanted content, and follow-for-follow accounts that seek to inflate follower counts artificially. The motivations behind the creation of these fake accounts are as varied as their forms, encompassing financial gain, social engineering, political influence, and even sheer mischief.

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_141

The application states the imposter profile creation on social networks is thought to be more harmful than any other type of cybercrime. It evaluates the effects of three supervised machine learning algorithms, Random Forest, Decision Tree, and Nave Bayes, supervised data mining algorithms (e.g., k-NN, decision tree, SVM, and naïve Bayes) in order to produce an appropriate prediction of fake or authentic profiles [1][9][5].

By combining image identification and natural language processing, the model uses machine learning to identify imposter Instagram accounts. Three machine learning classification algorithms—support vector machine (SVM), neural network (NN), and our recently developed technique, SVM-NN—were used to assess whether the target accounts were genuine or fraudulent. The results of both experiments showed that, in terms of accuracy rates, the supervised algorithms perform better than the unsupervised algorithms. [2][3]. In particular, the ID3 algorithm outperformed other classifiers in detecting phony Facebook profiles, and k-Medoids had the lowest detection accuracy rate. [10][6][15].

In the existing approach it is difficult to develop a system that can detect imposters with perfect accuracy and sometimes it falsely flags legitimate users as imposters. To overcome these disadvantages, new machine learning-based systems for profile imposter detection on Instagram should focus on ensemble learning technique which is the combination of machine learning algorithms. The application uses two machine learning algorithms which are XGBoost and Support vector machine which overcomes these types of allegations by improving the accuracy and fault tolerance of existing system. XGBoost provides high accuracy and trains data at high speed. It is more scalable and flexible. SVMs have good generalization performance. It can handle high-dimensional data without overfitting.

2.Literature Survey

Numerous efforts have previously been made in the area of Instagram automated and false account identification, which identifies automated and fraudulent accounts on the network using machine learning techniques including Naive Bayes, logistic regression, support vector machines, and neural networks. A cost-sensitive evolutionary algorithm is utilised to identify automated accounts due to the artificial bias in the dataset. For the automatic and fraudulent account detection issues, the results are 86% and 96%, respectively.[4].

There are two published datasets available for the purpose of identifying automated and phony accounts. Different machine learning techniques are used to detect these accounts. [7][11]. Instagram fake account identification by machine learning utilizing a combination of natural language processing and image detection. Using neural networks and Data Science is the most efficient way to identify phony accounts, according to a plethora of assessments, evaluations, and efforts [14][8].

In their research article, they suggested employing machine learning techniques, specifically logistic regression and random forest algorithm, to detect impostor accounts on Instagram in particular and attempt to determine whether an account is real or fraudulent. The Random Forest and Logistic Regression techniques yielded accuracy values of 92.5% and 90.8%, respectively [12]. This paper addresses the methodologies discussed above for detecting fraudulent social media profiles and clarifies the significance of fake identities in advanced persistent threats. Wewill evaluate the effectiveness of three supervised machine learning algorithms: Random Forest, Decision Tree, and Naive Bayes, in order to produce a pertinent prediction of phony or real profiles[13].As people using Instagram is increasing day by day there is vast growth in the imposter accounts with different tools like bots. So, no single machine learning algorithm is sufficient to provide accurate results.Fig.1 represents the algorithms used in existing model. In the existing approach it is difficult to develop a system that can detect imposters with perfect accuracy, and it falsely flags legitimate users as imposters

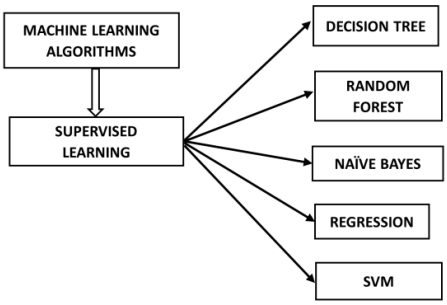

Fig. 1. Algorithms Used in existing models

3. Methodology

The suggested architecture for faker detection on Instagram which was mentioned on Fig.2 leverages the power of Support Vector Machines (SVM) and XGBoost computations to strengthen protections. In order to spot potential scams, this framework will analyze sample customer behavior, content, and commitment. Highlights such as consistency of posts, content, cooperation from those who follow, and log history will be removed. Clients will be sorted into trustworthy and questionable categories using the SVM algorithm, and XGBoost will help improve accuracy via ensemble learning. The system would provide risk rankings based on user actions compared to a database of legitimate users, enabling Instagram to take corrective measures, such as highlighting or suspending accounts with a high fake chance. This preventative approach will help to secure the platform's clientele and maintain its credibility.

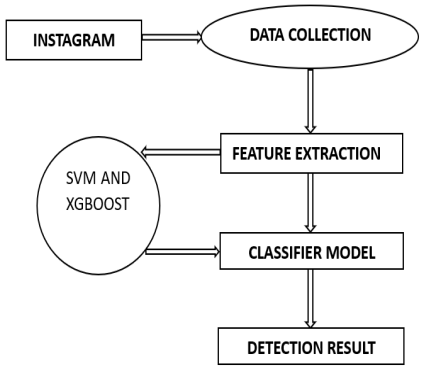

Fig. 2. Process flow diagram of the model

3.1 Data collection: Collecting a varied dataset of Instagram profiles, including real and fake accounts, is necessary for the false account identification on Instagram. To represent the "real" class in your dataset, gather a representative sample of actual Instagram profiles. These profiles are recognizable via legal methods. You can look for well-known fake accounts or search for phony accounts using other criteria, like suspicious behavior or activity. Make sure each account has a ground truth label indicating whether it is true or false. For supervised machine learning, this labeling procedure is essential.

3.2 Data Preprocessing

Instagram data can be collected from various sources like Instagram API, Instagram scrapping tools and Instagram analytical tools. So, the need to be preprocessed in order to perform analytics. In preprocessing it removes duplicate data, handles missing values and it converts data into consistent format.

3.3Training & Test Dataset: The train/test split is a critical step in machine learning and model evaluation. It involves dividing a dataset into two separate subsets:

Training Set: The machine learning model is trained using this section of the dataset. From this selection of data, the model discovers patterns and correlations within the data.

***Testing Set*:** The efficacy of the model is assessed using the testing set, which is maintained apart from the training set.. It provides an unbiased assessment of how well the trained model generalizes to new, unseen data.

The split is typically done randomly, with a common split ratio being 70-80% for training and 20-30% for testing, although the exact ratio may vary depending on the dataset's size and characteristics. The goal is to ensure that the model can make accurate predictions on data it has never encountered during training, helping to assess its real-world performance and identify potential issues like overfitting.

4.Proposed Methodology

Techniques for machine learning XGBoost and Support Vector Machine (SVM) are the foundations of the proposed system.

4.1 Support Vector Machine (SVM)

The SVM uses exceptional hyperplanes to divide all attributes of a particular sort in order to classify data as mentioned in Fig.3. The optimum hyperplane for a support vector machine (SVM) is the one where the lines connecting the classes are the longest. In order to categorize data, support vector machines (SVMs) look for the exceptional hyperplane that divides the different facets of knowledge into their corresponding categories. The support vectors are the informative features that are most similar to the keeping-apart vectors.

Implementation

Step 1: Open the dataset of Instagram users from sklearn. datasets
Step 2: Divide the goal variables and input characteristics.
Step 3: Train the SVM classifiers.
Step 4: Plot the input feature scatter plot.
Step 5: Plot the decision boundary.
Step 6: Plot the decision border in.

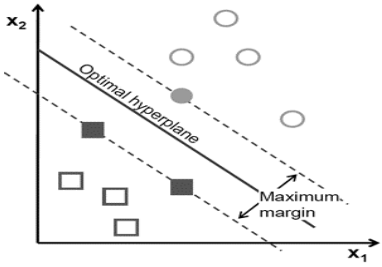

Fig.3 SVM Working

4.2 XGboost Algorithm

One of the quickest, most scalable unsupervised machine learning techniques for both regression and classification is called XGBoost. It combines the Number of weak in order to produce an exact result. One of the main features of XGBoost is its ability to handle big data sets and missing values successfully. The efficacy of the model is assessed using the testing set, which is maintained apart from the training set. Furthermore, XGBoost has built-in parallel processing capabilities that enable rapid model training on big datasets.Fig.4 explain the overall working view od XGBosst.

XGBoost Implementation

Step1: Reads the data set

Step2: Identifies the independent variables and the dependent variable.

Step3: Splits the data into training testing using a 70-30 ratio

Step4: Define the XGBoost model without any specific parameter tuning, i.e. leaving everything as de fault

Step5: Utilize the training data set to train the algorithm.

Step6: Apply the trained model to the testing data set

Step7: Obtain the accuracy

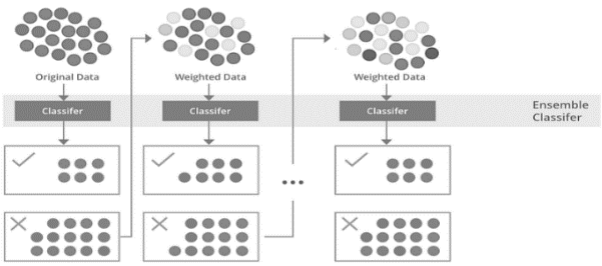

Fig. 4. XGBoost Working

5. Results and Discussions

Outline the measures that were used to assess the findings and choose the best supervised machine learning method.First demonstrate the model's precision before utilizing the confusion matrix from Formula 1. The following metrics are used to visualize the performance of the

various methods using this matrix. This indicator shows the percentage of cases that were returned that were legitimate phony profiles.

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN) \quad (1)$$

In this study, we analyzed the output of two machine learning algorithms (XGBoost and Support Vector Machine) to ascertain which method would work best for distinguishing real accounts from false ones in the Instagram dataset. The following tables(Table 1,Table 2) provide a summary of the final confusion matrix values for each algorithm based on the computation of successfully and incorrectly classified occurrences. It also displays the computation of accuracy for every algorithm. Predictions regarding algorithm performance indicate that XGBoost outperforms Support Vector Machine techniques with an accuracy score of 93.8%, it came in first. 90.5% is the result of the Support Vector Machine algorithm.The results is displayed on Fig.5 , whereas in existing model it only consists of 85-90% accuracy.

Table 1:Classification of profiles in instagram using XGBoost-Confusion matrix:

Confusion matrix:

	Predicted Class	**Not Predicted Class**
Original Account	810	15
Fake Account	10	120

Instances that are correctly classified:93.8%

Instances that are incorrectly classified :6.2%

Table 2: Classification of profiles in instagram using Support Vector Machine(SVM)

Confusion matrix:

	Predicted Class	**Not Predicted Class**
Original Account	800	20
Fake Account	10	120

Instances that are correctly classified:90.5%

Instances that are incorrectly classified:9.5%

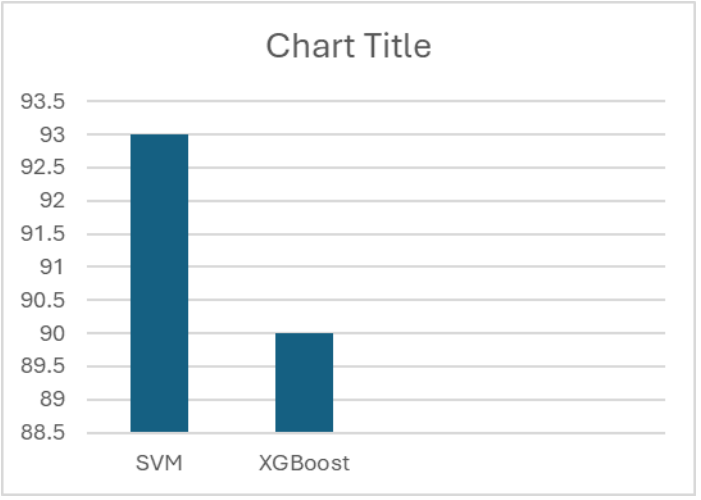

Fig 5:Graphical representation of Result

Table 3: Comparitive Study of the Approaches

	Existing Approaches	Proposed Approach
Algorithms	Decision Tree, Random Forest	SVM, XGBoost
Accuracy	85-90%	93-95%
Advantages	Simple to implement	High accuracy, More Scalable, Flexible
Disadvantages	Can be overfitting, Less Scalable	computationally expensive

5 Conclusion and Future Work

The current system discovered that not much research had been done specifically on Instagram as a social network platform while looking through earlier, comparable studies on the identification of phony profiles on social media platforms. For this reason, the suggested method used a targeted approach. In this research, the model presented a novel method based on machine learning techniques for identifying fake user accounts on Instagram based on specific features. For this, we employed the support vector machine and XGBoost models, respectively. before research for other social media networks has not yielded as high accuracy levels (the best accuracy reached before to this was 85%).

In order to identify photos on Instagram with better precision, we intend to rework current machine learning algorithms in the future by utilizing additional resources, like identifying items with neural systems. The suggested methodology not only increases accuracy but also boosts the effectiveness of current machine learning, enabling direct integration into social media software and doing away with the need for a separate program to identify phony accounts. The machine learning technique described in this study can be used with minor adjustments for LinkedIn and other social networking sites. With any lucky, this paper will contribute to future research efforts to decrease the amount of false accounts that are now on social media.

References

[1].I.Anupriya, V.Sowmiya, G.Devika, Prediction of fake instagram profiles using machine learning. Journal of Emerging Technologies and Innovative Research (JETIR).2023 March; Vol 10,Issue 3:1-6.

[2].Aljabri M, Aljameel SS, Mirza S, Intelligent techniques for detecting network attacks. Molecular Diversity Preservation International (MDPI).2021 Oct; Vol 21,1-43

[3]. Albayati M, Altamimi, machine learning model for detecting fake Facebook profiles using supervised and unsupervised mining techniques. International Journal of Simulation- Systems, Science and Technology- IJSSST.2019 Feb; V20.

[4]. Akyon FC, Esat Kalfaoglu M, Instagram fake and automated account detection. Innovations in intelligent systems and applications conference, ASYU.2022 May.

[5].Ananya Dey Hamsashree Reddy, Manjistha Dey, Niharika Sinha, Detection of fake accounts on Instagram using machine learning.International Journal of Computer Science & Information Technology (IJCSIT).Oct 2019;Vol 11. No 5.

[6] Kumar, Voruganti Naresh, U. Sivaji, Gunipati Kanishka, B. Rupa Devi, A. Suresh, K. Reddy Madhavi, and Syed Thouheed Ahmed. "A Framework For Tweet Classification And Analysis On Social Media Platform Using Federated Learning." Malaysian Journal of Computer Science (2023): 90-98. [7]. Krutika Palav, Pranali Awari, Instagram Fake Account Detection. International Research Journal of Modernization in Engineering Technology and Science.2021 May; Vol03, Issue:05.

[8]. Khushboo Saraswat, Nirupma Tiwari, Fake Account Detection in Social Media, International Journal Of Creative Research Thoughts (IJCRT).2021 Sept; Vol 10, Issue 09,Page 741-755.

[9]. S. P. Maniraj, Harie Krishnan G, Fake Account Detection using Machine Learning and Data Science, International Journal of Innovative Technology and Exploring Engineering (IJITEE) .2019 Nov; Vol9 Issue-1, Page 583-585.

[10]. Dr. Suchita Amey Bhovar, A Study of Different Methodologies to Detect Fake Account on Social Media using Machine Learning, International Journal of Science and Research (IJSR).2023 Feb;Vol 12, Issue 2,Page 53-57.

[11].Koosha Zarei, Reza Farahbakhsh, Noel Crespi,Typification of Impersonated Accounts on Instagram.2021 Oct; UTC from IEEE Xplore.

[12]. F. C. Akyon and M. Esat Kalfaoglu, Instagram Fake and Automated Account Detection. Innovations in Intelligent Systems and Applications Conference (ASYU).2019 Oct: 1-7.

[13]. Juandreas Ezarfelix, Nathanael Jeffrey, Novita Sari, A Systematic Literature Review: Instagram Fake Account Detection Based on Machine Learning. Binus Journal.2022 April; Vol. 4 No. 1

[14]. Yasyn Elyusufi, Zakaris Elyusufi, and Mohamed Ait Kbir, Social Network Fake Profile Detection Using Machine Learning Algorithm, Springer ;2020 Feb;pp.30-40.

[15]. Albayati MB, Altamimi AM, an empirical study for detecting fake Facebook profiles using supervised mining techniques. Slovenian Society Informatika.2019 March; Vol 3, No 1.