International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1995-2000
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0291

Detecting Deceptive Fake Profiles in Online Social Media

Naveenkumar N[1], Sridharshini B[2], Vijayakumar T[3], Thendral T[4], Puviyarasan A[5]
[1,2,3,4,5]Computer Science and Engineering Muthayammal Engineering College (Affiliated to Anna University.) Namakkal, Tamil Nadu, India.
Emails: *nnaveenkumar.dr@gmail.com[1], srivarshicse2004@gmail.com[2], viji321321@gmail.com[3], thendralt2003@gmail.com[4], puviashok404@gmail.com[5]*

Abstract
In Online Fake profiles present a security issue, online social networks (OSNs) help people communicate. Starting with data cleaning of the MIB dataset and manually collected data, this research use machine learning (ML) to identify fraudulent accounts. After testing a number of machine learning models, the Random Forest (RF) classifier showed the best accuracy. Max Voting (Majority Voting) increases cross-validation accuracy, whereas Extreme Gradient Boosting (XGBoost) and Decision Trees boost performance through ensemble learning. To ascertain validity, the algorithm examines profile details such as follower count, profile ID, and username. The RF classifier successfully identifies fake accounts, according to the results.
Keywords: *Social Networks, Machine Learning, Random Forest, XGBoost, Max Voting, Data Cleaning, Fake Profile Detection.*

1. Introduction
People can communicate globally through online social networks (OSNs), but a significant problem is the popularity of fake profiles. Fraudulent acts like identity theft, disinformation, and cybercrimes frequently use fake accounts The purpose of this project is for determining using machine learning methods. Fraud profile. The suggested technique attempts to categorize accounts as authentic or fake by examining a variety of profile characteristics and behavioral patterns. Improve Classification accuracy, machine learning model, etc. Random forests, XGBOOST trees are used. The main focus on the project is Identifying the bulk profiles in online social media. The traditional methods are used for identifying the single profile which is outdated. However, these social network spots to locate and hire devoted and skilled workers in different fields for their organizations

2. Methodology
2.1. Data Collection and Preprocessing
- The MIB dataset, containing real and fake accounts, was used for training and testing.
- Data preprocessing involved removing inconsistencies,

Handling missing values, and normalizing the dataset. The datasets were comprised of 3272 proper user accounts and 3351 fakes. The datasets also utilized TWT, INT, and FSF to test the fraudulent accounts. The learned information was kept in CSV form for machine extractions [1].

2.2. CSV File Conversion

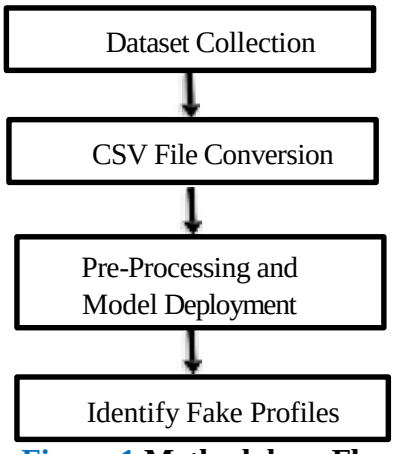

Figure 1 Methodology Flow

Due to the size of large data records in Microsoft Excel. Export these Excel files to CSV format so that this data can be used by planned software.

Steps to follow:

1. Choose File.
2. Choose “Save As” from the menu.
3. Name the file and select the .csv suffix (Comma melimited).

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1995-2000
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0291

4. Choose the Save option [9], shown in Figure 1.

2.3. Random Forest

Random forest also called random-decision forest, shows how ensemble learning works. Machine learning often uses this method because it works well for classifying and predicting things. Instead, it looks at what most of the trees predict to make its guess. But random forest makes way more decision trees than the usual method, and its answer comes from adding up what almost all these trees say [2]. Random forest helps spot profiles. The model takes in info and gives useful results. Tree (FB) corresponds to a sample Given Set 1 2, NX X X = and 1 2, NY Y = Response Early aggregation. It picks a random sample (B times) at set times [10]. How to calculate the given results Sample after training (x)

$$f = \frac{1}{B}\sum_{b-1}^{B} f_b(x') \qquad (1)$$

Random Forest can be expressed as:

y^=1N∑i=1NTi(x)\hat{y}=\frac{1}{N}\sum_{i=1}^{N} T_i(x)y^=N1i=1∑NTi(x) (2)

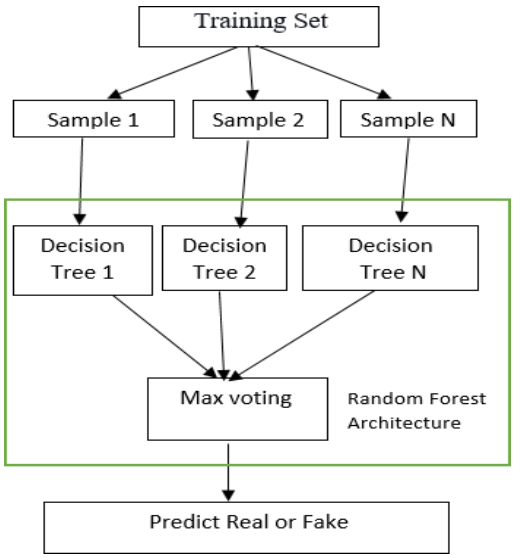

Figure 2 RF Workflow

2.4. Max Voting (Majority Voting)

This type of voting, also referred to as majority voting, is a method used in decision-making processes. Ensemble learning is a technique employed in classification problems to improve accuracy and reliability. By merging the forecasts of several machines, the works achieve a more accurate prediction. Learning models and choosing the class that will benefit from the highest number of votes as the final prediction, Figure 2. This technique enhances precision and minimizes the likelihood of errors. Misclassification can be achieved by utilizing the strengths of various factors models [3]. The Formula for Max Voting:

Final_Class = arg max (Count(Class_Predictions))

Advantages of Using Max Voting in This Project

- Higher Accuracy – By combining multiple models, the system reduces individual model errors.
- Robust Performance – The final classification is less affected by a single weak model.
- Better Generalization – Works well across different types of fake profiles.

2.5. Extreme Gradient Boost

XG Boost stands out as a unique method for ensemble learning in regression. He uses latent in various environments of probability of Gradient increase. Unlike random forest which shines when all data is present and accounted for, this method sidesteps the issue with a slick gradient-boosting tactic. The author jumps in with equation by setting up f0(x) as the starting point in the boosting game plan [4][5].

$$f_0(x) = \arg\min_{\gamma} \sum_{i-1}^{n} L(y_i, \gamma)$$

Here's how to calculate the results for a specific sample (x') after training.

$$\gamma_{im} = -\alpha\left[\frac{\delta L(y_i, F(x_i))}{F(x_i)}\right]$$

Mathematical Formula of XGBoost:

Obj= L(θ)+Ω(θ)Obj = L(\theta)\Omega(\theta)Obj= L(θ)+Ω(θ)

2.6. Decision Tree

A Decision Tree is this cool tool for guided learning. It's like a map you can use if you've got to sort things

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1995-2000
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0291

out or guess numbers, but it's super good at the sorting part. Imagine it like this big tree where each fork in the branches is a question about your data, the branches are choices you make, and every leaf on the tree is an answer you're looking for. In this project, the decision tree classifier is employed as a crucial machine learning model is being developed to detect fake content. Profiles in online social networks (OSN). It helps in understanding the decision-making process by analyzing different attributes of a social media profile and classifying it as real or fake based on predefined rules [6].

2.6.1. Selection Tree Terminologies

- **Root Node:** The basis node is from where the choice tree begins. It represents the whole dataset, which in addition gets divided into two or extra homogeneous units.
- **Leaf Knot:** Leaf Node -Last Output Node, and after receiving, the tree can no longer be separated. Leaf knot.
- **Splitting:** Splitting is the manner of dividing the selection node/root node into sub-nodes in line with the given situations.
- **Branch/Sub Tree:** A tree fashioned by splitting the tree.
- **Pruning:** Pruning is the technique of casting off the unwanted branches from the tree.
- **Parent/Child Node:** The basis node of the tree is called the discern node, and different nodes are called the kid nodes [7], Figure 3.

Formula: Expected value (EV) = (First possible outcome x Likelihood of outcome) + (Second possible outcome x Likelihood of outcome) – Cost.

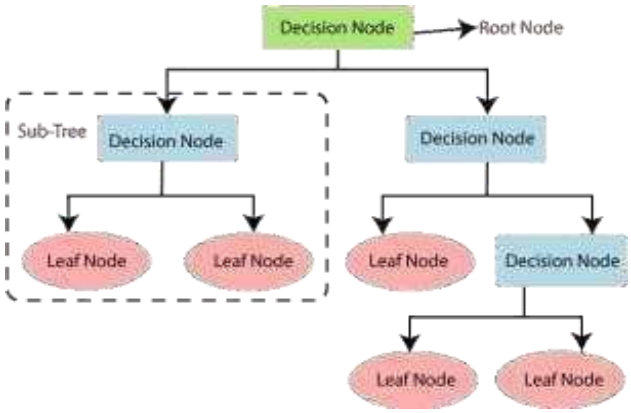

Figure 3 Decision Tree Flow

2.6.2. Attribute Selection Measures

When you set up a Decision tree, picking the top attribute for the root and the smaller branches gets tricky. But, this thing called the Attribute Selection Measure, or ASM for short saves the day. With ASM, choosing the right attribute for those tree bits becomes a piece of cake. Now, everyone raves about two big- deal techniques for ASM, which are:

- Information Gain
- Gini Index

2.6.3. Data Gain

- Data advantage is the measurement of adjustments in entropy after the segmentation of a dataset based on a characteristic.
- Information Gain= Entropy(S)- [(Weighted Avg) * Entropy (each feature)]

2.6.4. Gini Index

- The Gini index is a measure of contamination or purity used when creating a decision tree in a cart (classification and regression trees) [20].
- Gini Index= 1- ∑jPj2

2.6.5. Feature Selection for Splitting the Tree

- The Decision Tree splits the dataset based on features like profile activity, number of followers, and post frequency.
- The Gini Index calculates the impurity of each split and selects the best one.

A lower Gini Index value means a purer node (more instances of a single class, i.e., real or fake) [8], shown in Figure 4.

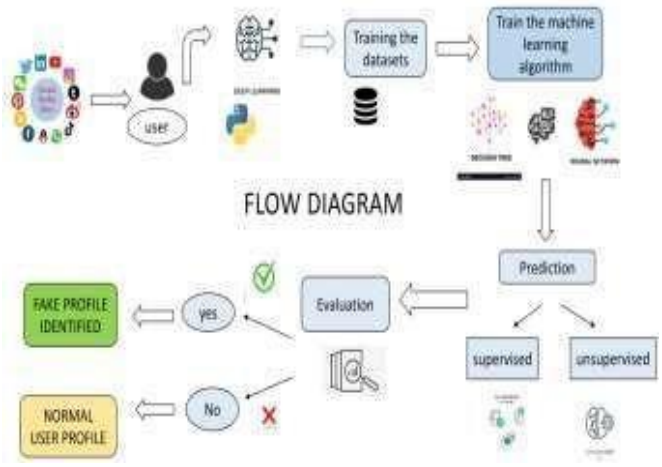

Figure 4 System Architecture

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1995-2000
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0291

- All tokens are written in lowercase, and stop words are no longer permitted in tweets.
- Then the word of this tweet with block chain support was built into the vector expression using the produced layer.
- All tokens are written in lowercase but no words. It is allowed in the long term in tweets. [9].

2.6.6. Outcomes of The Experiment

So here's what we're looking at: for the "LSTM neural network," we've got stuff like the Random Forest and Decision Tree, along with Regression and a few other strategies. Plus, these cool charts are showing the loss against epochs, and we can even check out how accurate the models were side by side. Then, we talk about what all this means and chew over the findings.

- Random Forest achieved the highest accuracy (98.3%), making it the most effective model for detecting fake profiles.
- XGBoost (94.6%) and Decision Trees (94.9%) performed well but had slightly lower accuracy than Random Forest.
- Max Voting (95.2%) improved classification stability, showing that an ensemble approach enhances model performance.
- Low False Positive and False Negative Rates, ensuring a high reliability in distinguishing between genuine and fraudulent profiles

3. Discussion

A fake account can change the same idea as influence. Popularity that can affect the national economy, Political system and social structure. They raise threats Application of social networks. As the authors explained in the introduction, this research uses several algorithms that can help identify fraudsters' profiles, which means that consumers will not be deceived or affected negatively by such malicious individuals. A blacklist was developed by the researchers of the previous study to effectively differentiate between fake attributes and fraudulent accounts. This study compares some things Machine learning algorithm showing machine learning algorithm Last results (xgboost 94.6%) The result of this method was higher than the first (94.9%) than that of former(91.1%). Deep Profile was made known as an approach that involves the application of a supervised learning algorithm to predict malicious accounts in a research that employed dynamic CNN Other relevant studies [10], other related studies Interesting approach was related to the identification of the sybil function. More based on registration hours. Phone number and address with Sybil. There was a false bet the positive aspects are 7%,3% and 21% in other cities. The authors of this study have achieved amazing 95% accuracy indicators. The method of sorting any forests and Woods was 98.3%of the synthetic profile predictions and used a virtual profile to use signs of signs [11], shown in Figure 5 & Table 1.

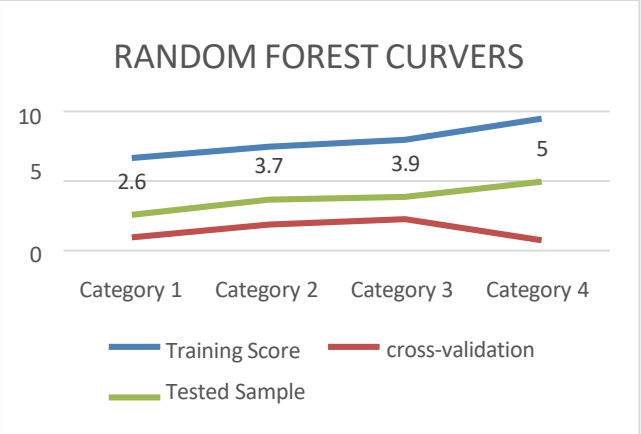

Figure 5 Random Forest (Learning Curves)

Table 1 Result Analysis for Fake Profiles

Metric	Random Forest	Decision Tree	XGBoost	Max Voting
Accuracy (%)	98.3	94.9	94.6	95.2
Precision (%)	96.8	92.4	93.1	94.0
Recall (%)	97.5	93.6	94.2	95.0
F1 Score (%)	97.1	93.0	93.6	94.5
False Positive Rate	Low	Medium	Low	Very Low
False Negative Rate	Very Low	High	Low	Medium

4. Confusion Matrix

A confusion matrix helps to visualize the result of a categorization problem by giving an arrangement of

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1995-2000
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0291

all outcomes of the prediction and findings in tabular form, shown in Figure 6 & Figure 7.

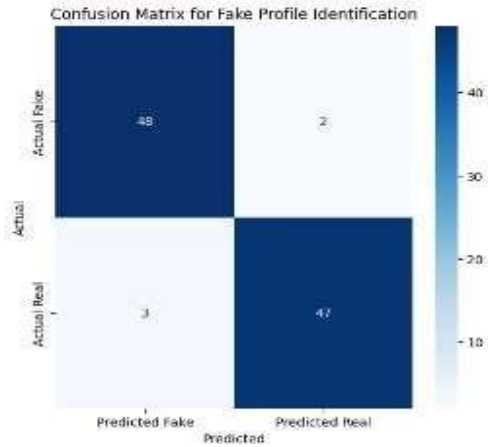

Figure 6 Confusion Matrix

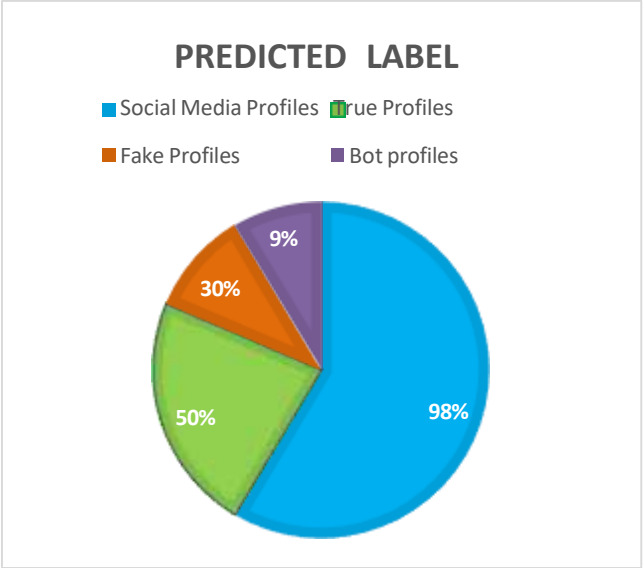

Figure 7 Predicted Label

Conclusion and Future Enhancement

They have used the easily accessible information for this architecture, which employed the CNN Model, Random Forests, and XG Boost supervised learning approaches in training the system on how to identify fake accounts. The main advantage of this project is to identify the bulk fake profiles in online social media. By analyzing the Decision tree, Random forest classifier and max voting techniques we can give 98% accuracy. While complicated, these hybrid models should produce better results. Future improvements could include integrating numerical, categorical, and profile image data into CNN models, introducing additional parameters, and combining multiple models to enhance accuracy. Developing a real-time detection system would further improve security on platforms like LinkedIn, Snapchat, WeChat, and QQ, making social networks safer and more reliable.

References

[1]. B. Erçahin, Ö. Aktaş, D. Kilinç , and C. Akyol, ‘‘Fakeaccount detection’’in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Oct. 2017, pp. 388–392.

[2]. Ferrara and Kudugunta, S. (2018) Bot detection using deep neural networks. 312–322 in Information Sciences, 467. https://doi.org/10.1016/j.ins.2018.08.019

[3]. Fake Profile Detection Methods in Large-ScaleOnline Social Networks: A Complete Study, D. Ramalingam and V. Chinnaiah, 2018.165–177 in Computers & Electrical Engineering, vol. 65. https://doi.org/10.1016/j.compeleceng.2017.05.020

[4]. K. Thomas, C. Grier, D. Song, and V. Paxson, ‘‘Suspended accounts inretrospect: An analysis of Twitter spam,’’ in Proc. ACM SIGCOMMConf. Internet Meas. Conf.,

[5]. M. Mazza, M. Avvenuti, S. Cresci, and M. Tesconi, ‘‘Investigating the difference between trolls, social bots, and humans on Twitter,’’ Comput.Commun., vol. 196, pp. 23–36, Dec. 2022.

[6]. Jie, H.J., and Wanda, P. (2020) DeepProfile: Utilizing Dynamic CNN to Detect Fake Profiles in Internet Social Networks.52, Article ID 102465 in Journal of Information Security and Applications. https://doi.org/10.1016/j.jisa.2020.102465

[7]. Identification of Fake Accounts on Twitter Using Hybrid RF Algorithm, Kodati, S., Reddy, K.P., Mekala, S., Murthy, P.S., and Reddy, P.C.S., 2021.Article No. 01046 of E3 S Web of Conferences, Page 309. https://doi.org/10.1051/e3sconf/202130901046

[8]. T. Wu, S. Wen, Y. Xiang, and W. Zhou,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 04 April 2025
Page No: 1995-2000
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0291

‘‘Twitter spam detection: Surveyof new approaches and comparative study,’’ Comput. Secur., vol. 76, pp. 265–284, Jul. 2018.

[9]. [R. C. Harkreader, and G. Gu, ‘‘Fake Account Dataset Detection’’in Recent Advances in Intrusion Detection, R. Sommer, D. Balzarotti, andG. Maier,

[10]. Eds. Berlin, Germany: Springer, 2018, pp. 318–337 [10]. YanjunQi., “Random forest for Bioinformatics in machinelearningalgorithm”.www.cs.cmu.edu /~qyj/papersA08/11-rfbook.pdf.

[11]. [Current Developments in Signal and Image Processing, edited by S. Bhattacharyya, L. Mri, M. Brkljai, J. V. Kureethara, and M. Koeppen, Springer, Singapore,95– 103. https://doi.org/10.1007/978-981-33-6966-5.