Detection of Fake Social Media Accounts Using Machine Learning

Mythri Chanda[1], Manasa Suresh[2], Hemanth N[3], C Lakshmi Preethi[4], Keerthana[5]
school Of Computer Science and Engineering, Reva University

***Abstract*—The proliferation of fake social media accounts poses a significant challenge to online security, trust, and authenticity. These fraudulent accounts are often used for malicious activities such as misinformation propagation, phishing, and spamming, necessitating robust detection mechanisms. This paper presents a machine learning-based approach for the detection of fake social media accounts using a dataset containing real and synthetic user profiles. The dataset is preprocessed using advanced techniques, including handling missing values, feature scaling, and text vectorization to enhance data quality.**

This research contributes to the ongoing efforts to enhance digital platform security by providing an efficient and scalable detection framework. The findings suggest that a hybrid approach combining rule-based and machine learning techniques can further improve detection accuracy. Future research will focus on integrating real-time detection mechanisms, leveraging graph-based anomaly detection, and expanding datasets to improve generalizability across multiple platforms and languages.

***Index Terms*—Fake accounts, Machine learning, social media security**

I. INTRODUCTION

Social media has revolutionized communication but has also introduced significant cybersecurity threats. One of the most concerning threats is the creation of fake accounts, which are used for activities such as misinformation spreading, phishing, and identity fraud. These fraudulent accounts undermine trust in digital platforms and can have severe consequences for users and businesses alike.

The rapid advancement of artificial intelligence (AI) and machine learning (ML) has provided an opportunity to develop automated detection methods that can identify fake accounts with high accuracy. Traditional rule-based systems struggle to keep up with the evolving tactics of malicious actors, making ML-driven solutions more effective.

The objective of this study is to design a machine learning framework that can efficiently detect fraudulent social media accounts. By leveraging different algorithms, feature engineering techniques, and evaluation metrics, we aim to improve detection accuracy and contribute to the security of online platforms

II. LITERATURE REVIEW

Online social networks (OSNs) have grown in popularity among today's youth, influencing their social life and motivating them to sign up for various social media platforms. Social media sites offer the required tools for a range of tasks, including news generation, Fake accounts have grown to be a serious issue with the growth of social media, endangering user security and platform integrity. In this work, we investigate how well machine learning (ML) algorithms identify phone accounts on social media sites like Twitter and Instagram. To train ML models for spotting fake accounts, we examine user behavior and account attributes, extracting parameters like the number of followers, activity level, and posting behavior. To preprocess the data and use different ML techniques, such Random Forest, Support Vector Machines, and XG boost, to categorize identify bogus accounts, we employ Python packages. The findings demonstrate that ML algorithms can accurately detect patterns and abnormalities suggestive of phone accounts and achieve high precision in fake account detection [1].

The proprietors of Fake Account extricate the individual data about others and spread the produced information on interpersonal organizations. This Paper mainly focuses on Fake Account detection, using classification techniques (Random Forest, K-Nearest Neighbor, Neural Network with (Sigmoid Activation Function)) from machine learning. Our proposed model can identify the fake account with the possible minimum set of attributes.[2]

This paper introduces an innovative approach for discerning and categorizing counterfeit social media profiles by leveraging the majority voting approach. The proposed methodology integrates a range of machine learning algorithms, including Decision Trees, Boost, Random Forest, Extra Trees, Logistic Regression, AdaBoost and K-Nearest Neighbors each tailored to capture distinct facets of user behavior and profile attributes. This amalgamation of diverse methods results in an ensemble of classifiers, which are subsequently subjected to a majority voting mechanism to render a conclusive judgment regarding the legitimacy of a given social media profile.[3]

The popularity of social media continues to grow, and its dominance of the entire world has become one of the aspects of modern life that cannot be ignored. The rapid growth of social media has resulted in the emergence of ecosystem problems. Hate speech, fraud, fake news, and a slew of other issues are becoming un-stoppable. With over 1.7 billion fake accounts on social media, the losses have already been significant, and removing these accounts will take a long time. Due to the growing number of Instagram users, the need for identifying fake accounts on social media, specifically in Instagram, is increasing. Because this process takes a long time if done manually by humans, we can now use machine learning to identify fake accounts thanks to the rapid development of machine learning. We can detect fake accounts on Instagram using machine learning by implementing the combination of image detection and natural language processing.[4]

On-Line social networks (OSNs) have become increasingly popular. People's social lives have become more associated with these sites. They use on-Line social networks (OSNs) to keep in touch with each other, share news, organize events, and even run their own e-business. The rapid growth of OSNs and the massive amount of personal data of its subscribers have attracted attackers, and imposters to steal personal data, share false news, and spread malicious activities. On the other hand, researchers have started to investigate efficient techniques to detect abnormal activities and fake accounts relying on accounts features, and classification algorithms. However, some of the account's exploited features have a negative contribution in the final results or have no impact, also using standalone classification algorithms does not always achieve satisfactory results. In this paper, a new algorithm, SVM-NN, is proposed to provide efficient detection for fake Twitter accounts and bots, feature selection and dimension reduction techniques were applied. Machine learning classification algorithms were used to decide the target account's identity real or fake, those algorithms were support vector machine (SVM), neural Network (NN), and our newly developed algorithm, SVM-NN. The proposed algorithm (SVM-NN) uses less features, while still being able to correctly classify about 98% of the accounts of our training dataset.[5]

Our lives are significantly impacted by social media platforms such as Facebook, Twitter, Instagram, LinkedIn, and others. People are actively participating in it all over. However, it also has to deal with the issue of bogus profiles. False accounts are frequently created by humans, bots, or computers. They are used to disseminate rumors and engage in illicit activities like identity theft and phishing. So, in this project, the author'll talk about a detection model that uses a variety of machine learning techniques to distinguish between fake and real Twitter profiles based on attributes like follower and friend counts, status updates, and more. The author used the dataset of Twitter profiles, separating real accounts into TFP and E13 and false accounts into INT, TWT, and FSF. Here, the author discusses LSTM, XG Boost, Random Forest, and Neural Networks. The key characteristics are chosen to assess a social media profile's authenticity. Hyperparameters and the architecture are also covered. Finally, results are produced after training the models. The output is therefore 0 for genuine profiles and 1 for false profiles. When a phony profile is discovered, it can be disabled or destroyed so

that cyber security problems can be prevented. Python and the necessary libraries, such as Sklearn, Numpy, and Pandas, are used for implementation. At the end of this study, the author will come to the conclusion that XG Boost is the best machine learning technique for finding fake profiles.[6]

III. METHODOLOGY

The described data pipeline for analysing social media data begins with Data Collection, specifically gathering information from social media platforms. This raw data then undergoes Data Preprocessing, where it is cleaned, formatted, noise is removed, missing values are handled, and formats are standardized for analysis. Following this, Data Validation ensures the pre-processed data meets quality standards; if not, an error is flagged, and the process returns to preprocessing. Once the data is validated, Feature Extraction identifies and extracts relevant features suitable for machine learning. The pipeline then diverges based on the learning paradigm chosen in Model Selection: Supervised Learning is employed with labelled data, while Unsupervised Learning is used to discover patterns in unlabelled data. In the Model Training/Clustering phase, a predictive model is trained in the supervised path, whereas data points are grouped into clusters in the unsupervised path. Subsequently, Evaluation/Pattern Identification occurs, where the supervised model's performance is assessed using relevant metrics, and patterns within the unsupervised clusters are identified and analysed. For the supervised path, a Performance Check determines if the model's performance is satisfactory; if so, it proceeds to Deployment; otherwise, parameters are tuned, and the model is re-evaluated. Finally, Monitoring continuously tracks the deployed model's performance, and Issue Detection triggers model retraining if problems are identified; if no issues arise, the process concludes. This architecture incorporates validation checkpoints, performance monitoring, and feedback loops, aligning with best practices for machine learning systems working with social media data to address its unique challenges and maintain model quality throughout its lifecycle**.**

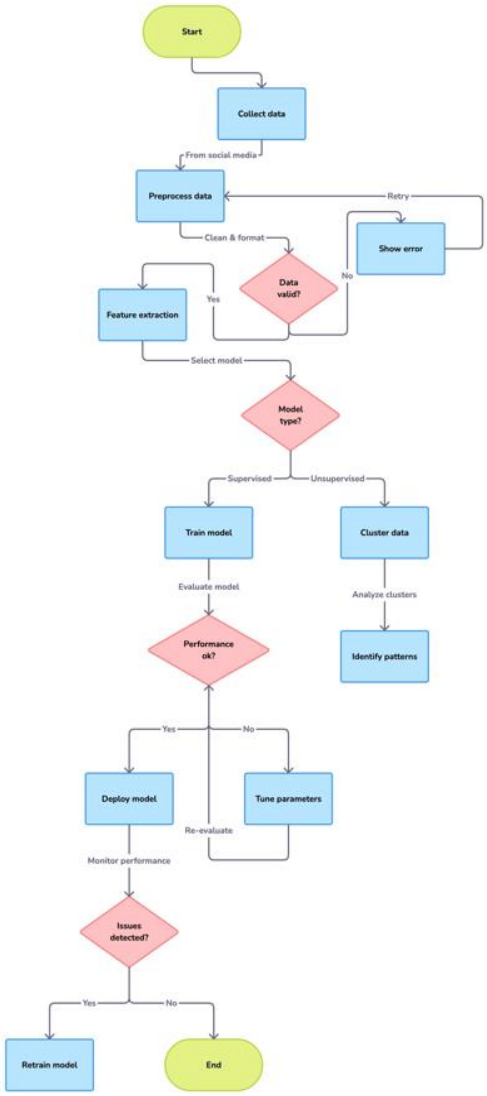

Fig. 1. System architecture for fake account detection using machine learning

IV. EVALUATION METRICS

To assess model performance, the following evaluation metrics are used:

Accuracy: The percentage of correctly classified instances.

Precision: The proportion of true positive detections among predicted positive cases.

Recall: The proportion of actual fake accounts that are correctly identified.

F1-score: The harmonic mean of precision and recall.

AUC-ROC: A graphical representation of a model's ability to distinguish between classes.

V. RESULTS AND DISCUSSIONS

Random Forest and Gradient Boosting achieved high precision and recall. ANN captured complex fraud patterns. Hybrid systems improved detection. Challenges include data imbalance and generalization. Model performance is evaluated using a confusion matrix to represent the classification results, as shown in Fig. 2.

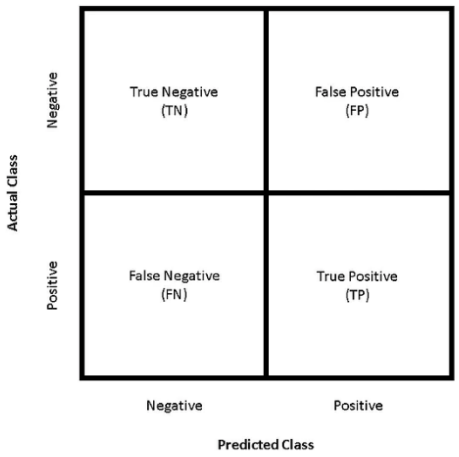

Fig. 2. Confusion matrix of the classification model for fake account detection

The ROC curve in Fig. 3 compares the true positive rate and false positive rate, providing insight into model sensitivity and specificity.

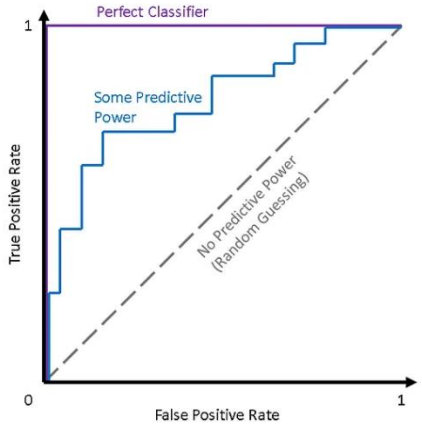

Fig. 3. ROC curve showing model performance trade-offs

To understand the relationships between various features used for detecting fake social media accounts, a Pearson correlation heatmap was generated. This helps identify which variables are strongly correlated and might influence the prediction results. For instance, the ratio of the number of letters in the username to the full name shows a moderate positive correlation, while follower and following counts show weak correlation to other features.

	ratio_numlen_username	len_fullname	ratio_numlen_fullname	len_desc	num_posts	num_followers	num_following
ratio_numlen_username	1.000000	-0.225472	0.408567	-0.321170	-0.157442	-0.062785	-0.172413
len_fullname	-0.225472	1.000000	-0.094348	0.272522	0.073350	0.033225	0.094855
ratio_numlen_fullname	0.408567	-0.094348	1.000000	-0.117521	-0.057716	-0.027035	-0.067971
len_desc	-0.321170	0.272522	-0.117521	1.000000	0.144824	0.005929	0.226561
num_posts	-0.157442	0.073350	-0.057716	0.144824	1.000000	0.321385	0.098225
num_followers	-0.062785	0.033225	-0.027035	0.005929	0.321385	1.000000	-0.011066
num_following	-0.172413	0.094855	-0.067971	0.226561	0.098225	-0.011066	1.000000

Figure 4: Correlation heatmap of user features used for fake account detection.

To evaluate classifier performance, the ROC (Receiver Operating Characteristic) curve is an effective visualization. The ROC curve compares the True Positive Rate (TPR) against the False Positive Rate (FPR). A classifier closer to the top-left corner performs better. This illustrative diagram shows how classifiers can be compared based on their curve positions relative to random and perfect classifiers.

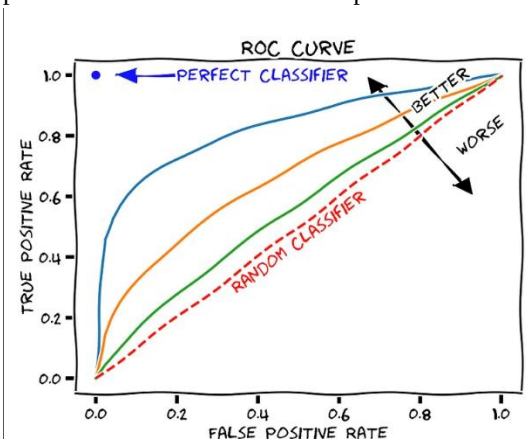

Figure 5: Conceptual ROC curve illustrating classifier performance levels.

The ROC curve plotted for the model trained in this study shows a strong performance with high recall and low false positive rate. Compared to the random and ideal models, the red line representing the trained model stays significantly above the baseline, indicating that the model has a strong ability to distinguish between real and fake accounts.

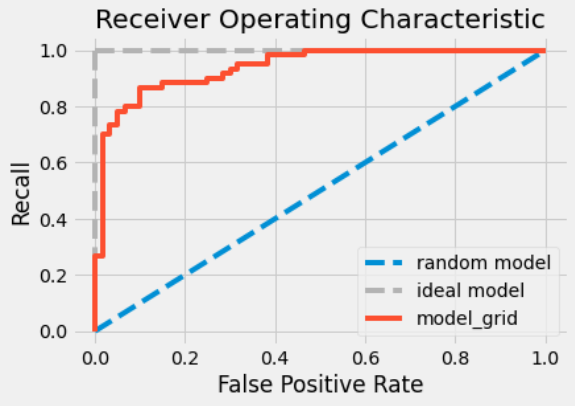

Figure 6: ROC curve comparing the trained model with random and ideal classifiers.

VII. CONCLUSION

This paper presents a machine learning-based fake account detection framework. Ensemble models performed best. Future work includes real-time analysis, graph-based detection, and multilingual dataset expansion.

VIII. ACKNOWLEDGMENT

The authors would like to thank Reva University for support and guidance throughout the project.

REFERENCES

[1] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi and M. Tesconi,"Faker vs. Faker: A Survey on Fake Accounts in Online Social Networks," IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2240–2271, 2015. doi: 10.1109/COMST.2015.2425944

[2] A. Kudugunta and E. Ferrara, "Deep Neural Networks for Bot Detection," in Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 1–8. doi: 10.1109/ASONAM.2018.8508646

[3] S. Kaur and A. Kaur, "Machine Learning Approach to Detect Fake Profiles on Social Media Platforms," in 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020, pp. 111–115. doi: 10.1109/ICIEM48762.2020.9160049

[4] S. Ahmed and M. Abulaish, "A Generic Statistical Approach for Spam Detection in Online Social Networks," Computer Communications, vol. 36, no. 10-11, pp. 1120–1129, 2013. doi: 10.1016/j.comcom.2013.03.004

[5] A. Fire, D. Goldschmidt and Y. Elovici, "Online Social Networks: Threats and Solutions," IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 2019–2036, 2014. doi: 10.1109/COMST.2014.2321628