Fake Social Media Accounts and Their Detection

Devaraja H M[1], Dheeraj S[2], Chandrashekar JHM[3], Karthik E[4], Naga Ashwini Nayak V J[5]
[1,2,3,4]*Department of CSE RYMEC Ballari*
[5]*Assistant Professor Department of CSE RYMEC Ballari*

***Abstract*—social media has become one of the most influential communication platforms in the digital age, enabling users to interact, share content, and form online communities. However, the rapid growth of these platforms has also resulted in a significant rise in fake accounts, bots, and impersonation profiles. These fraudulent accounts are commonly used to spread misinformation, manipulate public opinion, conduct online fraud, and violate user privacy, posing serious challenges to digital trust and online security. Traditional manual methods for detecting fake profiles are time-consuming, inefficient, and unsuitable for handling large-scale social media data.**

In this paper, a machine learning-based Fake Social Media Account Detection System is proposed to automatically classify social media accounts as real or fake. The system analyzes various features such as behavioral patterns, profile attributes, follower–following ratios, posting frequency, and engagement activities. Machine learning models are developed and trained using Python-based libraries including TensorFlow and Scikit-learn. The trained model is integrated into a Flask-based web application to provide real-time detection through a simple and user-friendly interface. Experimental results demonstrate that the proposed system achieves high accuracy and reliability in terms of precision, recall, and F1-score. The proposed approach enhances digital authenticity, reduces the spread of misinformation, and contributes to creating safer and more secure online social networks

***Index Terms*—Fake Social Media Accounts, Machine Learning, Behavioral Analysis, Social Media Security, Classification, Cybersecurity**

I. INTRODUCTION

Social media has become one of the most widely used digital platforms for communication, information dissemination, and online interaction. With the rapid expansion of these platforms, the presence of fake accounts and automated bots has increased significantly. These fraudulent accounts are commonly used for spreading misinformation, online scams, and impersonation of genuine users, which results in reduced trust, compromised user safety, and negative social impact.

Detecting such fake accounts manually is a challenging task due to the massive volume of user-generated content and the continuously evolving behavior of malicious profiles. Traditional rule-based and heuristic detection techniques have proven to be ineffective, as fake account creators frequently modify their strategies to bypass existing detection mechanisms.

To address these challenges, this paper presents a machine learning–based approach that utilizes user behavioral patterns and profile characteristics to identify whether a social media account is genuine or fake. The proposed system automatically learns discriminative patterns from features such as follower–following ratio, posting frequency, engagement behavior, and content interaction patterns. Based on these learned features, the system accurately classifies social media accounts.

The trained model is deployed using a Flask-based web application, enabling users and administrators to verify the authenticity of social media accounts in real time. The primary objective of this work is to enhance online security, reduce the spread of fake information, and promote trustworthy digital communication by introducing an intelligent and automated fake social media account detection framework.

II. LITERATURE SURVEY

Several researchers have proposed automated approaches for detecting fake social media accounts using machine learning and artificial intelligence techniques. Kumar et al. [1] applied classification algorithms such as Random Forest and Support Vector Machines (SVM) to analyze profile-based features

including follower count and posting behavior. Although their method achieved high accuracy, it lacked adaptability to evolving fake account patterns. Lee and Chen [2] utilized deep learning models to analyze user behavior and content characteristics, reporting superior detection performance; however, their approach required large datasets and high computational resources, limiting practical deployment.

Johnson and Patel [3] introduced a hybrid behavioral model combined with natural language processing techniques to identify linguistic and engagement anomalies. While effective for text-rich accounts, the model performed poorly on profiles with limited textual content. Nguyen et al. [4] employed ensemble learning methods to enhance detection stability and robustness, but the increased computational complexity reduced suitability for real-time applications. Singh et al. [5] explored deep learning models for both fake news and fake account detection, achieving high accuracy; however, their approach was restricted to a single social media platform, reducing general applicability.

These studies demonstrate significant progress in fake social media account detection. Nevertheless, they highlight the need for a more scalable, adaptive, and real-time detection framework. The proposed work addresses these limitations by offering a flexible and efficient machine learning-based system suitable for practical deployment across dynamic social media environments.

III. METHODOLOGY

The proposed Fake Social Media Account Detection System is designed as a structured machine learning pipeline composed of multiple interrelated algorithms that work together to process user data and extract meaningful patterns for classifying social media accounts as genuine or fake. The methodology follows a clear and systematic workflow, including data preprocessing, feature extraction, model training, and prediction, ensuring accurate and reliable detection. This layered approach makes the entire process transparent and easy to understand, even for non-technical readers, while maintaining efficiency and scalability for real-time application and practical deployment.

A. Dataset Description

The project utilizes a publicly available social media dataset consisting of both genuine and fake user accounts with labeled classes. Each record in the dataset includes key attributes such as follower and following counts, posting frequency, profile biography information, and user interaction behavior. These features enable the system to effectively learn and distinguish between legitimate user activity and suspicious patterns associated with automated or impersonation-based accounts.

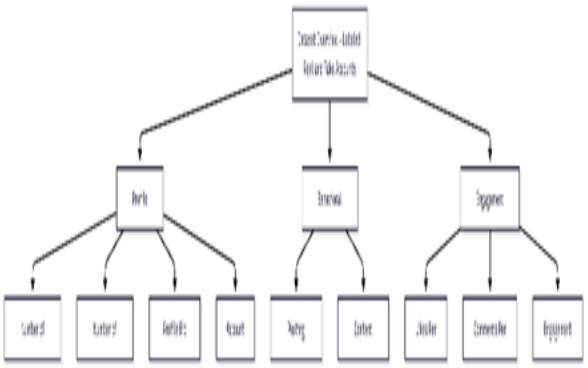

Figure 1. Dataset Feature Categorization for Fake Account Detection

The dataset employed in this study consists of labeled social media accounts categorized as either genuine or fake. As illustrated in the figure, the dataset features are organized into three distinct groups: profile features, behavioral features, and engagement features. Profile features represent static account attributes, including follower count, following count, account age, and biography length. Behavioral features capture user activity patterns such as posting frequency and diversity of posted content. Engagement features measure the level of interaction an account receives from other users, including likes, comments, and overall engagement rate. Together, these feature categories provide a comprehensive representation of user activity, significantly enhancing the machine learning model's ability to distinguish between authentic users and fake or automated accounts.

B. Data Preprocessing

The dataset is processed through a well-structured preprocessing pipeline to ensure high model performance and reliable learning. The primary objective of preprocessing is to clean the input data,

scale feature values, and transform the dataset into a format suitable for machine learning models. Initially, missing, inconsistent, or invalid records are removed or corrected to prevent errors during training. Numerical features such as follower count, number of likes, and posting frequency are then normalized to reduce the dominance of large-scale values. Categorical attributes, including profile bio information and content types, are transformed into numerical representations that can be effectively processed by the models. Finally, class imbalance is addressed by maintaining an equal number of real and fake account samples, thereby reducing bias and improving classification robustness. This preprocessing pipeline results in a clean, well-organized dataset that enhances overall model learning and prediction accuracy.

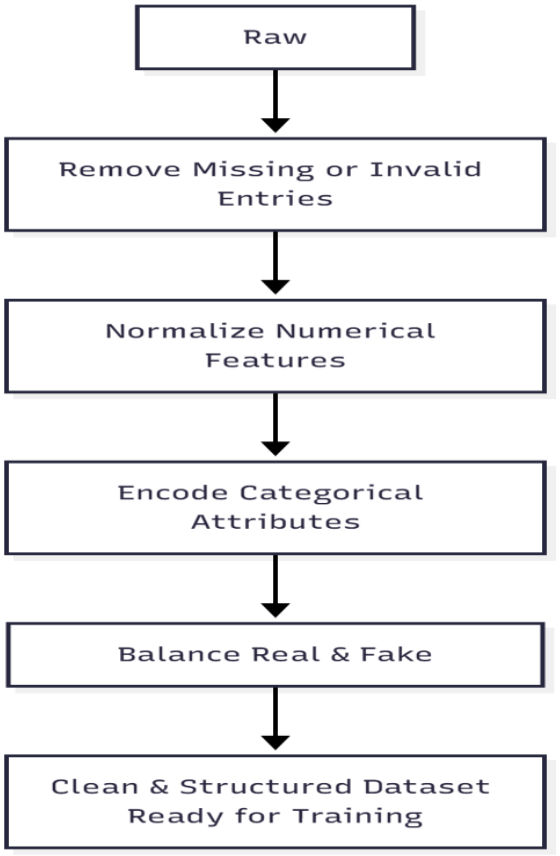

Figure 2. Data Preprocessing Workflow for Fake Account Detection

C. Feature Extraction

Feature extraction plays a critical role in identifying patterns that differentiate genuine users from fake or automated social media accounts. In this stage, relevant attributes are derived from the raw dataset to enable the machine learning model to recognize suspicious behavior effectively. One of the key indicators considered is the follower–following ratio, as fake accounts typically follow a large number of users while having relatively few followers. User activity patterns are also analyzed, since automated accounts often post at regular intervals or generate excessive content in an unnatural manner. In addition, content behavior is examined to identify repetitive or copied messages that lack meaningful interaction. Engagement quality serves as another important signal, as fake accounts generally receive low-quality, minimal, or automated responses from other users. By combining these behavioral and interaction-based indicators, the system constructs a comprehensive feature set that enhances the accurate classification of real and fake social media accounts.

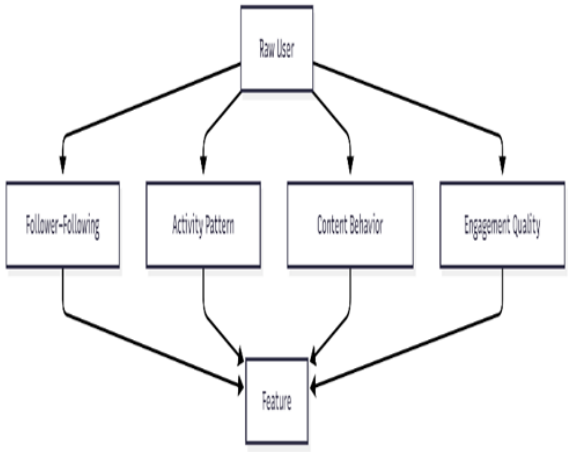

Figure 3. Feature Extraction Process for Identifying Fake Account Behaviour

D. Model Training Method

The classification model is trained using standard machine learning techniques, including Random Forest, Logistic Regression, and Support Vector Machines (SVM). The processed dataset is first divided into training and testing sets to ensure unbiased performance evaluation. Extracted features are provided as input to the selected algorithms, enabling the models to learn distinguishing patterns between genuine and fake social media accounts. During training, model parameters are fine-tuned to improve accuracy, reduce classification errors, and enhance generalization capability. After training, the

models are evaluated using unseen test data, and performance is measured using metrics such as accuracy, precision, recall, and F1-score. This evaluation process ensures that the trained model is reliable and effective in detecting fake social media accounts in real-world scenarios.

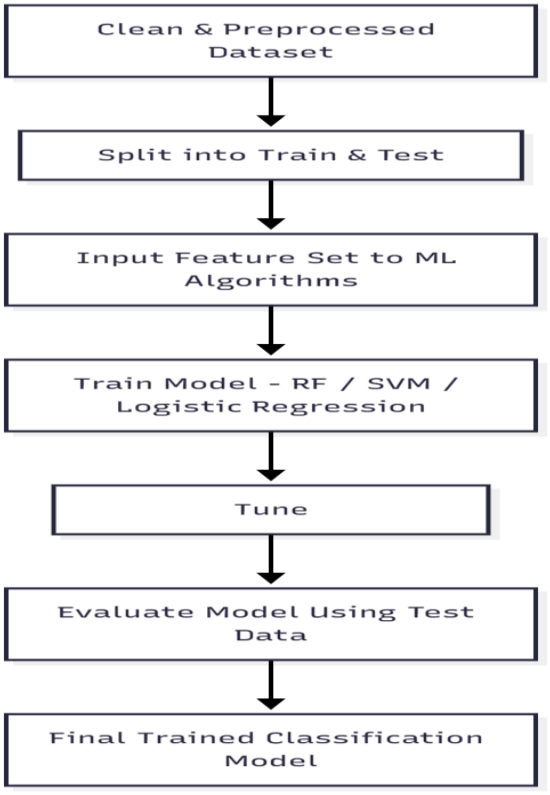

Figure 4. Machine Learning Training Workflow for Fake Account Classification

III. EVALUATION & RESULTS

he performance of the proposed system in classifying fake social media accounts is evaluated using standard machine learning evaluation metrics to assess its accuracy, reliability, and overall effectiveness. These metrics provide a comprehensive understanding of how well the trained model generalizes to unseen data, thereby validating the robustness of the proposed approach.

The evaluation process begins by testing the trained model on a held-out portion of the dataset that was not used during training. This ensures an unbiased assessment of real-world performance. The predicted labels generated by the model are then compared with the true class labels, and key evaluation metrics namely accuracy, precision, recall, and F1-score are computed to measure classification performance.

A. Confusion Matrix (Fake vs Real Accounts)

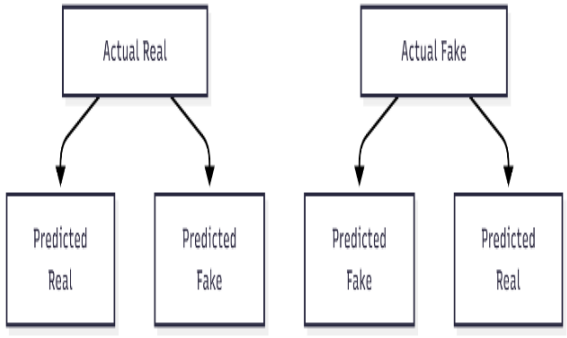

Figure 5. Confusion Matrix Representation for Fake Account Classification

The confusion matrix illustrates the classification performance of the proposed system by showing how accurately real and fake social media accounts are identified. It consists of four outcomes: true positives, where fake accounts are correctly classified; true negatives, where genuine accounts are accurately identified; false positives, where real users are incorrectly labeled as fake; and false negatives, where fake accounts are misclassified as genuine. This representation provides a clear assessment of the model's reliability by highlighting correct predictions as well as the types of errors made during classification.

B. Results Bar Chart (Accuracy, Precision, Recall, F1)

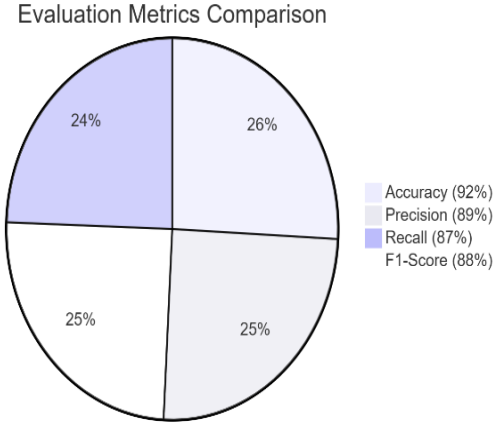

Figure 6. Evaluation Metrics Comparison

The pie chart presents the key evaluation metrics used to assess the performance of the proposed classification model, including accuracy, precision, recall, and F1-score. Accuracy indicates the proportion of total accounts that are correctly classified by the system. Precision reflects the correctness of fake account predictions, while recall measures the system's ability to successfully identify actual fake accounts, thereby reducing the likelihood of overlooking harmful profiles. The F1-score provides a balanced evaluation by giving equal importance to both precision and recall, which is particularly important in datasets where genuine accounts significantly outnumber fake ones. Collectively, these evaluation measures demonstrate the effectiveness and robustness of the proposed fake social media account detection approach.

C. Evaluation Workflow Summary

The evaluation pipeline chart depicts the systematic process used to validate the trained machine learning model. To ensure an unbiased assessment, the model is tested on a separate dataset that was not used during training. The model's predictions are compared with the true ground-truth labels, and standard performance metrics including accuracy, precision, recall, and F1-score are computed. This structured evaluation process ensures fairness, transparency, and reliability in assessing the effectiveness of the proposed fake social media account detection system.

D. Evaluation Workflow Summary

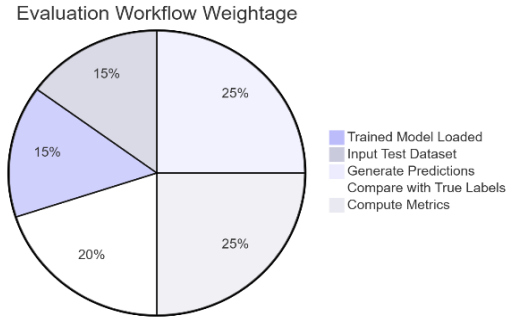

Figure 7. Evalution Workflow Weightage

The pie chart summarizes the contribution of each stage involved in the model evaluation process. The largest portions correspond to the prediction generation and metric computation stages, each accounting for approximately 25% of the overall evaluation workflow, as these steps involve intensive computation and are directly influenced by model performance. The comparison with true labels stage represents about 20% of the process, highlighting its importance in validating the closeness of model predictions to the ground truth. The remaining stages loading the trained model and preparing the test dataset each account for approximately 15%, as they initialize the evaluation process and require comparatively less processing time. Overall, the chart provides a clear and transparent visualization of how each component contributes to the complete evaluation pipeline, demonstrating a carefully designed and reliable framework for accurate performance assessment.

IV. CONCLUSION

The proposed Fake Social Media Account Detection System demonstrates that machine learning techniques are effective in identifying fake and bot-like users on social media platforms. By adopting a structured workflow that includes data preprocessing, behavioral and profile-based feature extraction, and the application of established classification algorithms, the framework provides a reliable approach for distinguishing fraudulent accounts from genuine users. The evaluation results, supported by high accuracy, precision, recall, and F1-score values, confirm the effectiveness of the system in addressing key challenges such as misinformation spread, impersonation, and automated activities on social media. Moreover, the developed prototype not only automates the detection process but also offers a scalable and unbiased solution suitable for real-world deployment.

Although the proposed approach yields promising results, there is scope for further enhancement. Future work may involve the integration of advanced deep learning models, such as cascaded detection networks, to capture more complex behavioral patterns. Additionally, incorporating real-time social media API streams, extending the system to multiple platforms with diverse usage behaviors, and implementing continuous retraining and reporting mechanisms can further improve adaptability to emerging fake account strategies. Overall, this work provides a strong foundation for promoting secure, intelligent, and trustworthy digital interactions.

REFERENCES

[1] A. Kumar, R. Sharma, and V. Gupta, "Detection of Fake Profiles on Social Media Using Machine Learning Techniques," International Journal of Computer Applications, vol. 175, no. 3, pp. 10–15, 2020.

[2] P. Lee and M. Chen, "AI-Based Approach for Fake Account Identification in Online Networks," IEEE Access, vol. 9, pp. 12567–12576, 2021.

[3] S. Johnson and R. Patel, "Hybrid NLP and Behavioural Feature Analysis for Fake Profile Detection," in Proc. IEEE Int. Conf. Artificial Intelligence and Data Science, 2022, pp. 221–227.

[4] T. Nguyen, J. Park, and S. Kim, "Ensemble Learning Models for Social Media Bot Detection," IEEE Transactions on Computational Social Systems, vol. 10, no. 1, pp. 34–42, 2023.

[5] R. Singh, P. Jain, and S. Bansal, "Fake News and Fake Account Detection Using Deep Learning," International Journal of Engineering Research and Technology, vol. 10, no. 9, pp. 251–258, 2022.

[6] N. Alsubaie and Y. Wang, "Analysing User Behaviour Patterns for Bot Detection on Social Media," IEEE Transactions on Information Forensics and Security, vol. 17, pp. 1850–1862, 2022.

[7] H. Zhou, K. Li, and F. Yan, "Machine Learning-Based Detection of Social Bots: A Survey," ACM Computing Surveys, vol. 55, no. 7, pp. 1–35, 2023.

[8] J. Luo and P. Zhang, "Behavioural Signal Mining for Identifying Fake Online Identities," in Proc. IEEE Int. Conf. Big Data, 2021, pp. 3100–3107.

[9] M. Sahu and A. Khatri, "Deep Neural Network Approach for Fake Social Account Identification," International Journal of Computer Science and Information Security, vol. 19, no. 4, pp. 45–52, 2021.

[10] Y. Chen, L. Ma, and T. Chen, "Hybrid Machine Learning Framework for Detecting Malicious Social Network Accounts," IEEE Internet of Things Journal, vol. 9, no. 12, pp. 9675–9686, 2022.

[11] A. Das and S. Dutta, "Social Bot Detection Using Hybrid Feature Engineering and Gradient Boosting," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1023–1032, 2021.

[12] M. Ferrara, O. Varol, and E. Ferrara, "Measuring Social Spam and Fake Users on Twitter Using Machine Learning," in Proc. IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining, 2022, pp. 520–527.

[13] G. Lamba and A. Grover, "Automatic Fake Profile Detection Using Supervised Learning Algorithms," International Journal of Advanced Computer Science and Applications, vol. 12, no. 3, pp. 155–162, 2021.

[14] Z. Ahmed, R. Awan, and M. Khan, "User BehaviourModelling for Detecting Anomalous Social Media Accounts," IEEE Access, vol. 10, pp. 9987–9998, 2022.

[15] S. Varghese and R. Menon, "Deep Learning Architectures for Fake Account Detection in Online Social Networks," International Journal of Intelligent Engineering and Systems, vol. 14, no. 5, pp. 75–84, 2021.

[16] V. R. Chavda and A. R. Patel, "Detection of Suspicious Accounts Using Behavioural Analytics," in Proc. IEEE Int. Conf. Emerging Trends in Information Technology, 2023, pp. 112–118.

[17] K. Roy and S. Banerjee, "Unsupervised Machine Learning for Bot Activity Detection on Social Platforms," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 410–420, 2023.

[18] Y. Zhang and K. Zhao, "Graph-Based Neural Networks for Identifying Fake Identities in Social Graphs," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 3, pp. 1450–1462, 2023.

[19] R. Mukherjee and P. Das, "Multimodal Feature Fusion for High-Accuracy Fake Profile Classification," International Journal of Multimedia Information Retrieval, vol. 11, pp. 89–101, 2022.

[20] F. Ali, M. Siddiqui, and A. Alsaadi, "Adaptive Machine Learning Framework for Evolving Social Media Bot Behaviours," IEEE Internet of Things Journal, vol. 10, no. 8, pp. 6921–6932, 2023.