International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Unmasking Fake Profiles - Machine Learning in Social Network Security

L.Sankara Rao[1], S.Manideep[2*], S.Dally[3] , M.V.N.Sai Saketh[4], T.Rohit Rao[5]

[1]Asst.Professor, Dept of CSE AI-ML, Raghu Institute of Technology
[2]Student, Dept of CSE AI-ML, Raghu Institute of Technology*
[3]Student, Dept of CSE AI-ML, Raghu Institute of Technology
[4]Student, Dept of CSE AI-ML, Raghu Institute of Technology
[5]student, Dept of CSE AI-ML, Raghu Institute of Technology

Abstract - The proliferation of fake profiles in social media poses significant threats to online security, privacy, and trust. Existing methods for detecting fake profiles rely heavily on manual verification, rule-based systems, or shallow machine-learning models, which are often ineffective against sophisticated fake profiles. This study proposes a novel machine learning framework for unmasking fake profiles in social media. Our approach leverages a combination of Extracting comprehensive features from user profiles, including behavioral, network, and content-based attributes, Data augmentation and Ensemble learning. Traditional machine-learning methods have limitations in detecting fake human accounts on social media. These accounts often mimic real users, making them difficult to distinguish. To overcome this challenge, we propose a new model. This new model can have more accuracy than the previous methods. By training the model on a dataset of real and fake accounts, we can improve the accuracy of detecting fake human accounts and enhance the security of social media platforms.

Key Words: Fake Profiles, Social Media, Instagram, Machine Learning(ML).

1.INTRODUCTION

Online social networks (OSNs) have evolved into a foundational element of contemporary society, linking billions of users worldwide and revolutionizing how people communicate, collaborate, and exchange information. Despite their transformative impact, the proliferation of counterfeit profiles has introduced severe challenges, such as diminished trust, data exploitation, and breaches of privacy. These fabricated accounts frequently mimic real individuals or entities, facilitating harmful activities like disseminating false information, swaying public sentiment, and executing cyberattacks. Such risks undermine the credibility and safety of OSNs, necessitating robust solutions to identify and counteract fraudulent profiles.

Machine learning (ML) has gained prominence as a critical tool for automating fake profile detection, leveraging its ability to process vast datasets and analyze behavioral patterns.To increase the accuracy of the model we implement several Algorithms.

Logistic Regression: A linear model for binary classification that uses a logistic function to model the probability of class membership.

DecisionTree: A tree-based model that splits the dataset on feature values to form branches leading to class labels.

RandomForest: An ensemble of multiple decision trees that improves predictive accuracy and reduces overfitting by averaging their results.

K-Nearest Neighbors: A non-parametric model that classifies a data point based on the labels of its nearest neighbors in the feature space.

ArtificialNeuralNetworks: A computational model consisting of interconnected layers of neurons that learns complex patterns from data through training.

So, we are including all these machine learning algorithms and select better one which gives best analysis.This helps in getting results more precise and scalable fake profile identification.

2. LITERATURE REVIEW

Literature Review on Unmasking Fake Profiles Using Machine Learning and Deep Learning Techniques.

The detection of fake profiles on social media platforms has become a critical area of research due to the increasing prevalence of fraudulent accounts and their associated risks. Various studies have explored machine learning (ML) and deep learning (DL) techniques to address this challenge, each contributing unique insights and methodologies.

[1] Focuses on identifying human-created fraudulent accounts using Random Forest with TF-IDF vectorization. The study achieved an accuracy of 87%, demonstrating the effectiveness of combining Random Forest with text-based feature extraction techniques for detecting fake profiles.

[2] Compares three popular algorithms—Logistic Regression, Decision Tree, and Random Forest—for fake account detection. The study concludes that the Random Forest classifier outperforms the others, making it a suitable choice for further research in this domain.

[3] Proposes a hybrid approach combining a Random Forest classifier with a Recurrent Neural Network (RNN) and a Logistic Regression meta-classifier. This method captures intricate relationships in structured user profile data and textual input, achieving promising results as evidenced by evaluation metrics such as accuracy scores and confusion matrices.

[4] Employs a Support Vector Machine (SVM) for binary classification in large datasets. Despite the non-linear decision boundary, the model achieves over 90% accuracy in distinguishing between fake and genuine profiles, highlighting the robustness of SVM for this task.

[5] Explores the use of Recurrent Neural Networks (RNNs) for detecting identity deception on social media. The proposed RNN-based classification method improves accuracy and precision compared to traditional machine learning algorithms, offering faster and more reliable results.

[6] Evaluates several supervised machine learning algorithms, including Logistic Regression, Bernoulli Naive Bayes, Random Forest, Support Vector Machine, and Artificial Neural Networks (ANNs), for detecting fake Instagram accounts. The Random Forest algorithm achieves the highest accuracy of 92%, making it the most effective model in this study.

[7] Investigates the use of both machine learning and deep learning techniques to detect fraudulent accounts. The study emphasizes the importance of ML in efficient data extraction and suggests extending the work to real-time social media data using APIs for future research.

[8] Introduces an Artificial Neural Network (ANN)-based model integrated into machine learning frameworks for fake account detection. The study anticipates improved precision and recall rates compared to traditional Random Forest-based systems, marking a significant advancement in the field.

[9] Leverages Artificial Neural Networks (ANNs) to automate the detection of fake Instagram accounts, eliminating the need for manual prediction. The study highlights the efficiency of ANNs in processing new test data to identify fake profiles, reducing the reliance on human resources and time-consuming processes.

[10] This project proposes a machine learning approach to detect fake Instagram profiles using SVM, KNN, Random Forest, Naive Bayes, and XGBoost. Random Forest demonstrated the highest accuracy in classification, with fraudulent profile IDs stored in a data dictionary for authorities. The work can be extended to address click log fraud, aiding cybercrime investigations.

[11] This paper evaluates various machine learning models for detecting fake profiles in OSNs, including SVM, KNN, Random Forest, MLP, Logistic Regression, and Naive Bayes. Random Forest achieved the highest accuracy (96%), while Naive Bayes had the lowest (74%). The study suggests that deep learning could further enhance detection.

[12] This paper explores machine learning techniques for detecting fake Instagram profiles, achieving 92.5% accuracy. While the results show promise, continuous refinement is needed to counter evolving fraudulent tactics. The study highlights the importance of vigilance, innovation, and collaboration in ensuring online community integrity.

In summary, these studies collectively highlight the effectiveness of machine learning and deep learning techniques, particularly Random Forest, SVM, and ANNs, in detecting fake profiles. Hybrid approaches combining multiple algorithms, such as Random Forest with RNNs, have shown promise in capturing complex patterns and improving detection accuracy. Future research directions include leveraging real-time data and further enhancing ANN-based models for greater precision and scalability.

3. METHODOLOGY

[3.1] Planning and Requirement Gathering

The first phase of the project involves understanding the problem, defining objectives, and gathering requirements. The primary goal of this project is to develop an Machine Learning-based system capable of detecting fake profiles on Instagram. Fake profiles pose significant risks, including scams, phishing, and misinformation, making it essential to identify and mitigate them effectively. To achieve this, the project focuses on collecting and analyzing Instagram profile data, extracting relevant features, and training an ML model to classify profiles as real or fake. The functional requirements include data collection, preprocessing, feature extraction, model training, and evaluation. Non-functional requirements emphasize scalability, accuracy, and usability. Tools and technologies selected for the project include Python for programming, TensorFlow/Keras for building the ANN, Pandas for data processing, and Scikit-learn for model evaluation. The dataset will consist of Instagram profile data, including both real and fake profiles, to ensure the model is trained on balanced and representative data.

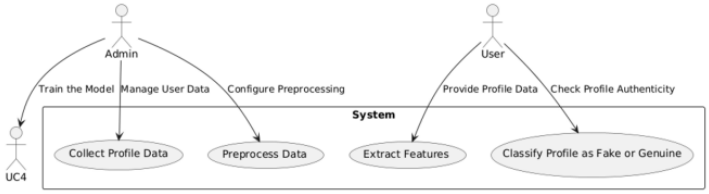

Fig-1: Architecture and Workflow.

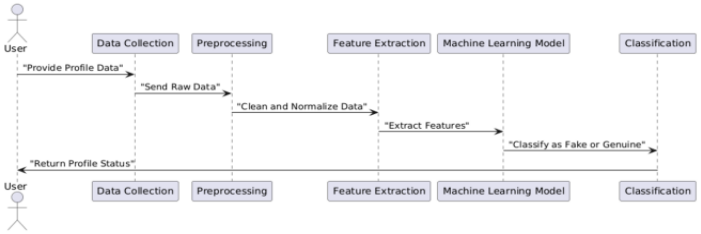

Fig-2 : Machine Learning Pipeline for Fake Profile Detection in Online Social Networks.

[3.2] Dataset Selection

The second phase involves creating a dataset suitable for training and testing the ML model. The dataset is a critical component of the project, as the model's performance depends heavily on the quality and relevance of the data. So, we created our own datasets using various attributes from various datasets. The dataset should include features such as Profile pic, username , username length, full name, full name length, name and user name relation, description length, external urls, private or public, posts, followers, follows, etc... This division ensures the model is trained on a substantial portion of the data while being validated and tested on unseen data to evaluate its generalization capability.

	profile pic	nums/length username	fullname words	nums/length fullname	name==username	description length	external URL	private	#posts	#followers	#follows	fake
0	1	0.27	0	0.00	0	53	0	0	32	1000	955	0
1	1	0.00	2	0.00	0	44	0	0	286	2740	533	0
2	1	0.10	2	0.00	0	0	0	1	13	159	98	0
3	1	0.00	1	0.00	0	82	0	0	679	414	651	0
4	1	0.00	2	0.00	0	0	0	1	6	151	126	0
...	...	...	...	...	...	...	...	...	...	...	...	...
571	1	0.55	1	0.44	0	0	0	0	33	166	596	1
572	1	0.38	1	0.33	0	21	0	0	44	66	75	1
573	1	0.57	2	0.00	0	0	0	0	4	96	339	1
574	1	0.57	1	0.00	0	11	0	0	0	57	73	1
575	1	0.27	1	0.00	0	0	0	0	2	150	487	1

576 rows × 12 columns

Table-1 : Dataset.

[3.3] Pre-processing

The third phase focuses on cleaning and preparing the dataset for training the ML model. Data preprocessing is a crucial step to ensure the dataset is free from inconsistencies and ready for analysis. The first step involves data cleaning, where missing

values are handled by filling them with mean/median values or removing rows/columns with excessive missing data. Duplicate profiles are also removed to avoid bias in the model.

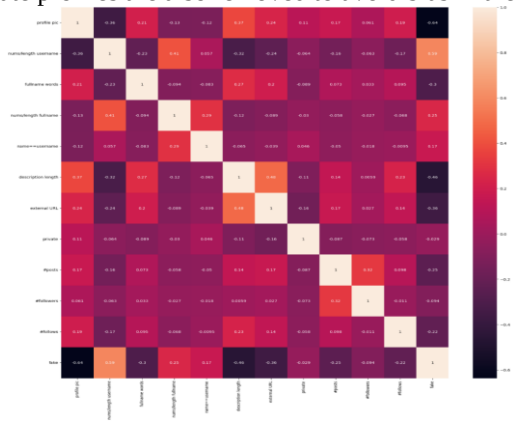

Fig-3 : Correlation between all features in datasets.

Next, feature engineering is performed to extract relevant features that can help distinguish fake profiles from real ones. Examples include analyzing username patterns (e.g., length, special characters), calculating the follower-to-following ratio, and performing image analysis on profile pictures to detect default or stolen images.

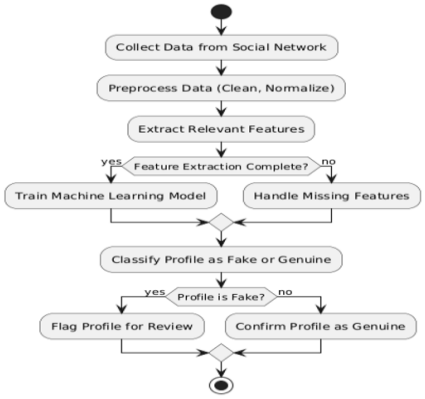

Fig-4 : Flow-diagram of preprocessing.

Numerical features are normalized or standardized to ensure they are on a similar scale, while categorical features are encoded using techniques like one-hot encoding. If the dataset is imbalanced, data augmentation techniques such as Synthetic Minority Over-sampling Technique can be applied to generate synthetic data for the minority class (fake profiles). Finally, the preprocessed dataset is split into training, validation, and test sets to facilitate model training and evaluation. This step ensures the model is trained on high-quality data and can generalize well to unseen profiles.

[3.4] Algorithms

Logistic Regression: This algorithm predicts the probability of an account being fake using features like follower-following ratios, post frequency, or account age. By applying a logistic function, it classifies accounts based on linear relationships between these features. It's efficient for binary outcomes but may struggle with complex, non-linear patterns in sophisticated fake accounts.

Decision Tree:A Decision Tree splits data using rules (e.g., profile completeness, activity spikes) to classify accounts. It's interpretable, showing how features like erratic posting times or generic bios flag fakes. However, it risks overfitting if overly complex, capturing noise instead of genuine patterns, which can reduce accuracy on unseen data.

Random Forest:This ensemble method combines multiple decision trees to improve accuracy and reduce overfitting. By aggregating predictions from trees trained on varied features (e.g., geolocation inconsistencies, bot-like comments), it handles complex interactions better than single trees. It's robust against noise, making it reliable for detecting evolving fake account tactics.

Artificial Neural Networks:ANNs detect intricate patterns through layered nodes, modeling non-linear relationships in data like image metadata or interaction networks. They excel with large datasets, identifying subtle cues (e.g., AI-generated content) but require significant computational resources and tuning. Deep learning variants can adapt to sophisticated fake behaviors.

K-Nearest Neighbors:K-NN classifies accounts by comparing them to labeled examples (e.g., fake accounts with sparse posts or repetitive comments). It flags accounts similar to known fakes in feature space. While simple, its performance depends on feature scaling and dataset size, and it may lag with real-time detection due to computational overhead.

4. RESULTS

Finally, Our Machine Learning model was trained and tested with Logistic regression, Decision Tree, Random forest, Artificial Neural Networks, K-Nearest Neighbour. From all these algorithms Logistic Regression and Random Forest performed same with accuracy of 91.Also ANN performed with least accuracy of 82.

Analysis of Logistic Regression: This classification report shows how well the Logistic Regression model can differentiate between fake (1) and genuine (0) Instagram profiles. The model achieves an overall accuracy of 0.91, meaning it correctly classifies 91% of accounts in the test set. The precision, recall, and F1-score for both classes indicate robust performance. Class 0 has slightly higher recall (0.95) but slightly lower precision (0.90), while class 1 shows higher precision (0.94) but lower recall (0.87). The macro and weighted averages reflect balanced results across classes, highlighting the model's capability to detect fake profiles effectively, while maintaining strong performance on genuine accounts overall.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.95      0.92        63
           1       0.94      0.87      0.90        53

    accuracy                           0.91       116
   macro avg       0.92      0.91      0.91       116
weighted avg       0.92      0.91      0.91       116

Accuracy Score: 0.9137931034482759
```

Fig-5: Classification report of Logistic Regression Model.

Analysis of Decision Tree: This classification report shows the performance of a Decision Tree model for identifying fake Instagram accounts, with class 0 representing genuine profiles and class 1 representing fake profiles. The overall accuracy is 0.87, indicating that 87% of profiles were correctly classified. Class 0 has a precision of 0.86, recall of 0.90, and F1-score of 0.88, meaning it effectively identifies genuine profiles while occasionally misclassifying them. Class 1, representing fake profiles, achieves a precision of 0.88, recall of 0.83, and F1-score of 0.85, showing reasonable detection ability. These results highlight the Decision Tree's balanced performance but leave room for further improvement.

Classification Report:

	precision	recall	f1-score	support
0	0.86	0.90	0.88	63
1	0.88	0.83	0.85	53
accuracy			0.87	116
macro avg	0.87	0.87	0.87	116
weighted avg	0.87	0.87	0.87	116

Accuracy Score: 0.8706896551724138

Fig-6: Classification report of Decision Tree Model.

Analysis of Random Forest: This classification report showcases how effectively the Random Forest model distinguishes between genuine (0) and fake (1) Instagram profiles. The model achieves an overall accuracy of about 91%, indicating that it correctly identifies the majority of accounts. With a precision of 0.94 for class 1, the model is highly confident in labeling accounts as fake. Meanwhile, its recall of 0.87 for class 1 means it correctly catches 87% of actual fake accounts. Class 0 shows a slightly higher recall of 0.95, ensuring genuine profiles are rarely misclassified. These results underscore the Random Forest's strong performance in robustly detecting fake accounts.

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.95	0.92	63
1	0.94	0.87	0.90	53
accuracy			0.91	116
macro avg	0.92	0.91	0.91	116
weighted avg	0.92	0.91	0.91	116

Accuracy Score: 0.9137931034482759

Fig-7: Classification report of Random Forest Model.

Analysis of Artificial Neural Networks: This classification report evaluates the performance of an Artificial Neural Network for detecting fake Instagram profiles, with class 0 representing genuine accounts and class 1 representing fake accounts. The model achieves an overall accuracy of 0.82, indicating that it correctly classifies most profiles. Class 0 has a precision of 0.90 but a recall of 0.75, meaning it accurately identifies genuine accounts but sometimes misses them. Conversely, class 1 shows lower precision (0.75) but higher recall (0.91), indicating a strong ability to catch fake accounts yet occasionally labeling genuine ones as fake. These results highlight the ANN model's overall balanced performance.

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.75	0.82	63
1	0.75	0.91	0.82	53
accuracy			0.82	116
macro avg	0.83	0.83	0.82	116
weighted avg	0.83	0.82	0.82	116

Accuracy Score: 0.8189655172413793

Fig-8: Classification report of ANN Model.

Analysis of K-Nearest Neighbour: This classification report shows the performance of a K-Nearest Neighbors model for detecting fake Instagram profiles, with class 0 representing genuine accounts and class 1 representing fake ones. The model achieves an overall accuracy of 0.86, meaning it correctly identifies 86% of the accounts. The precision, recall, and F1-scores for both classes are relatively balanced, indicating consistent classification across genuine and fake profiles. Class 0's precision and recall are 0.87 each, while class 1's are 0.85 each. This suggests the model is generally good at differentiating genuine from fake accounts, though there is still room for improvement in borderline cases.

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.87	0.87	63
1	0.85	0.85	0.85	53
accuracy			0.86	116
macro avg	0.86	0.86	0.86	116
weighted avg	0.86	0.86	0.86	116

Accuracy Score: 0.8620689655172413

Fig-9: Classification report of KNN Model.

5. CONCLUSIONS

Our project, "Unmasking Fake Profiles: Machine Learning in Social Network Security," successfully developed an Machine Learning-based system to detect fake Instagram profiles.Our Model shows the profile if real or fake with an accuracy of 91%,which is improved from previous models.

Among the five classification models employed to Unmask fake Instagram profiles—Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and Artificial Neural Networks—Random Forest and Logistic Regression demonstrated the highest overall accuracy, reaching around 91%. The Decision Tree model followed closely at 87%, while KNN and ANN achieved 86% and 82% respectively. Each algorithm showed varying strengths in precision and recall, reflecting trade-offs in capturing genuine and fake accounts. Overall, the ensemble-based Random Forest offered robust performance, whereas simpler models like Logistic Regression excelled with fewer parameters. Future enhancements could involve hyperparameter tuning, feature engineering, and larger datasets to boost accuracy.This project demonstrates a scalable, ethical solution for enhancing Instagram security, offering actionable insights for real-world deployment to combat misinformation, phishing, and fraud.

6. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who contributed to the successful completion of this project on Unmasking Fake Profiles - Machine Learning in Social Network Security.

First and foremost, we extend our deepest appreciation to our mentor L Sankara Rao(Assistant Professor), for their invaluable guidance, insightful suggestions, and continuous support throughout the research and development process. Their expertise and encouragement have played a significant role in shaping this project.

We would also like to thank our institution, Raghu Institute of Technology, and the project co-ordinator Dr.S Vidya Sagar for providing us with the necessary resources, infrastructure, and technical support. Their encouragement and constructive feedback have been instrumental in refining our approach.

Furthermore, we acknowledge the contributions of our peers, friends, and colleagues for their constant motivation and support. Their discussions and feedback have helped us overcome various challenges during the implementation phase.

Lastly, we express our gratitude to our families for their unwavering support and patience throughout this journey. Without their encouragement, this project would not have been possible.

7. REFERENCES

[1] K.Umbrani, D. Shah, A. Pile and A. Jain, "Fake Profile Detection Using Machine Learning," 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), Manama, Bahrain, 2024, pp. 966-973, doi: 10.1109/ICETSIS61505.2024.10459570.

[2] K. V. Nikhitha, K. Bhavya and D. U. Nandini, "Fake Account Detection on Social Media using Random Forest Classifier," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 806-811, doi: 10.1109/ICICCS56967.2023.10142841.

[3] S. Bhatia and M. Sharma, "Deep Learning Technique to Detect Fake Accounts on Social Media," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2024, pp. 1-5, doi: 10.1109/ICRITO61523.2024.10522400.

[4] S. R. Ramya, R. Priyanka, S. S. Priya, M. Srinivashini and A. Yasodha, "SVM Based Fake Account Sign-In Detection," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 509-514, doi: 10.1109/ICOEI56765.2023.10125850.

[5] B. S. Borkar, D. R. Patil, A. V. Markad and M. Sharma, "Real or Fake Identity Deception of Social Media Accounts using Recurrent Neural Network," 2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP), Uttarakhand, India, 2022, pp. 80-84, doi: 10.1109/ICFIRTP56122.2022.10059430.

[6] M. J. Ekosputra, A. Susanto, F. Haryanto and D. Suhartono, "Supervised Machine Learning Algorithms to Detect Instagram Fake Accounts," 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2021, pp. 396-400, doi: 10.1109/ISRITI54043.2021.9702833.

[7] K. Ramani, S. Bhargavi, S. Kowsar, N. Chowdary and M. Chennakesava, "Fake Account Detection Using Machine and Deep Techniques in Social Media," 2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 2024, pp. 706-710, doi: 10.1109/ICOECA62351.2024.00128.

[8] Y. A. Rani, G. Srividya, A. Balaram, K. H. Kumar, A. Kiran and M. Silparaj, "Fake Account Detection Using ANN Based Model in Machine Learning," 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India, 2024, pp. 1-7, doi: 10.1109/ICSTEM61137.2024.10561061.

[9] A. S. Fathima, S. Reema and S. T. Ahmed, "ANN Based Fake Profile Detection and Categorization Using Premetric Paradigms On Instagram," 2023 Innovations in Power and Advanced Computing Technologies (i-PACT), Kuala Lumpur, Malaysia, 2023, pp. 1-6, doi: 10.1109/i-PACT58649.2023.10434755.

[10] P. Harris, J. Gojal, R. Chitra and S. Anithra, "Fake Instagram Profile Identification and Classification using Machine Learning," 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-5, doi: 10.1109/GCAT52182.2021.9587858.

[11] J. Joseph and V. S, "Fake Profile Detection in Online Social Networks Using Machine Learning Models," 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE), Kerala, India, 2023, pp. 1-5, doi: 10.1109/RASSE60029.2023.10363482.

[12] R. R. Arunprakaash and R. Nathiya, "Leveraging Machine Learning algorithms for Fake Profile Detection on Instagram," 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2024, pp. 869-876, doi: 10.1109/ICCPCT61902.2024.10673398.