IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Multimodal Fusion Framework For Fake Social Media Profile Detection Using Machine Learning

Poluri lakshmi Prasanna
M.Tech., Scholar,
Department of CSE
QIS College of Engineering & Technology, Ongole, India

Dr. M. Senthil
Professor,
Department of CSE
QIS College of Engineering& Technology,
Ongole, India

Dr. Nidamanuri Srinu
Associate Professor,
Department of CSE
QIS College of Engineering& Technology,
Ongole, India

Abstract: At present social network sites are part of the life for most of the people. Every day several people are creating their profiles on the social network platforms and they are interacting with others independent of the user's location and time. The social network sites not only providing advantages to the users and also provide security issues to the users as well their information. To analyze, who are encouraging threats in social network we need to classify the social networks profiles of the users. From the classification, we can get the genuine profiles and fake profiles on the social networks. Traditionally, we have different classification methods for detecting the fake profiles on the social networks. But we need to improve the accuracy rate of the fake profile detection in the social networks. In this paper we are proposing Machine learning and Natural language Processing (NLP) techniques to improve the accuracy rate of the fake profiles detection. We can use the Support Vector Machine (SVM) and Naïve Bayes algorithm.

Keywords: Fake Profile Detection, Multimodal Fusion, Machine Learning, Social Media Security, Deep Learning, Meta-Classifier.

I. INTRODUCTION

Social networking has end up a well-known recreation within the web at present, attracting hundreds of thousands of users, spending billions of minutes on such services. Online Social network (OSN) services variety from social interactionsbased platforms similar to Instagram or Facebook or MySpace, to understanding dissemination-centric platforms reminiscent of twitter or Google Buzz, to social interaction characteristic brought to present systems such as Flicker. The opposite hand, enhancing security concerns and protecting the OSN privateness still signify a most important bottleneck and viewed

mission. When making use of Social Network's (SN's), one of a kind men and women share one-of-a-kind quantities of their private understanding. Having our individual knowhow entirely or in part uncovered to the general public, makes us excellent targets for unique types of assaults, the worst of which could be identification theft. Identity theft happens when any individual uses character's expertise for a private attain or purpose. During the earlier years, online identification theft has been a primary problem considering it affected millions of people's worldwide. Victims of identification theft may suffer unique types of penalties; for illustration, they would lose time/cash, get dispatched to reformatory, get their public image ruined, or have their relationships with associates and loved ones damaged. At present, the vast majority of SN's does no longer verifies ordinary users‟ debts and has very susceptible privateness and safety policies. In fact, most SN's applications default their settings to minimal privateness; and consequently, SN's became a best platform for fraud and abuse. Social Networking offerings have facilitated identity theft and Impersonation attacks for serious as good as naive attackers. To make things worse, users are required to furnish correct understanding to set up an account in Social Networking web sites.

With the rise of fake accounts and bots, it has become increasingly challenging to distinguish between real and fake accounts. These fake accounts can be used for various malicious purposes, such as spreading misinformation, phishing, and identity theft. In this paper, we will discuss a machine learning-based approach for identifying fake social media accounts. Our proposed method involves a multi-step process that combines various features to accurately identify fake accounts. The first step involves data collection and preprocessing. We will collect a large dataset of social media profiles, both real and fake, from various platforms such as Facebook, Twitter, and Instagram. The data will be cleaned and preprocessed to remove any irrelevant information and prepare it for further analysis.The second step involves feature extraction. We will extract various features from the preprocessed data, such as user behavior, network structure, content analysis, and account metadata. These features will be used to train our machine learning models.The third step involves model selection and training. We will experiment with different machine learning algorithms such as Support vector machines(SVM), KNearest Neighbors Algorithm (KNN), Random forest,Logistic Regression & Artificial Neural Network (ANN) to find the best-performing model for our task. The models will be trained on the preprocessed data and evaluated using various metrics such as accuracy, precision, recall, and F1 score.Once we have selected the best-performing model, we will deploy it on a production environment to identify fake accounts in real-time. We will also continuously monitor the performance of the model and fine-tune it as needed to improve its accuracy over time.Our proposed method for identifying fake social media accounts using machine learning is a multi-step process that combines data collection, feature engineering, model selection and training, and model deployment and evaluation. By leveraging the power of machine learning algorithms, we can accurately distinguish between real and fake social media accounts and mitigate the negative impacts of fake accounts on social media platforms.

II. RELATED WORKS

While the rapid advancement of social media has transformed the landscape for communication, it has also resulted in a number of security issues, with the creation of fake accounts standing out as a major threat that enables identity theft, disinformation, and online fraud, and thus requires automated detection systems. Current research can generally be grouped into one of two camps: traditional machine learning (ML) methods using engineered features, and more recent deep learning (DL) methods using multimodal data.

Fake profiles are used in advanced persistent threats and are also used in other nefarious activities. As we all know, Globally, billions of individuals utilize Social networking sites like Facebook, Twitter, LinkedIn, Instagram, etc. to establish connections. A new era of networking has been ushered in by social networks simplicity and accessibility. At the same time, various types of scammers are drawn to these social media platforms. These scammers make fake profiles to spread their content and carry out scams. In this project, we used Deep Neural Networking and Machine Learning algorithms namely Artificial Neural Networks(ANN), Random Forest and Support vector machine(SVM) algorithms to assess the likelihood that Facebook account information is accurate or not. The dataset used in this paper is taken from GitHub which is a Facebook profile Dataset to identify faux and genuine profiles, also we have described the associated classes and libraries. Here we are going to predict the faux and real profiles using the best accurate model after comparing the outcomes of the three techniques employed. [3]

Nazir et al. (2010) describes recognizing and describing phantom profiles in online social gaming applications. The article analyses a Facebook application, the online game "Fighters club", known to provide incentives and gaming advantage to those users who invite their peers into the game. The authors contend that by giving such impetuses the game motivates its players to make fake profiles. By presenting

those fake profiles into the game, the user would increase a motivating force of an incentive for him/herself. [4]

Adikari and Dutta (2014) depict recognizable proof of fake profiles on LinkedIn. The paper demonstrates that fake profiles can be recognized with 84% exactness and 2.44% false negative, utilizing constrained profile information as input. Techniques, for example, neural networks, SVMs, and Principal component analysis are applied. Among others, highlights, for example, the number of languages spoken, training, abilities, suggestions, interests, and awards are utilized. Qualities of profiles, known to be fake, posted on uncommon sites are utilized as a ground truth. [5]

Chu et al. (2010) go for separating Twitter accounts operated by humans, bots, or cyborgs (i.e., bots and people working in concert). As a part of the detection problem formulation, the Identification of spamming records is acknowledged with the assistance of an Orthogonal Sparse Bigram (OSB) text classifier that uses pairs of words as features. [6]

Stringhini et al. (2013) analyze Twitter supporter markets. They describe the qualities of Twitter devotee advertises and group the clients of the business sectors. The authors argue that there are two major kinds of accounts who pursue the "client": fake accounts("sybils"), and compromised accounts, proprietors of which don't presume that their followers rundown is expanding. Clients of adherent markets might be famous people or legislators, meaning to give the appearance of having a bigger fan base, or might be cybercriminals, going for making their record look progressively authentic, so they can rapidly spread malware what's more, spam. [7]

In 2019, Faiza Masood, Ghana Ammad, Ahmad Almogren, Assad Abbas, Hasan Ali Khathak, Ikram Uddin,MohsenGuizani, and MansourZuair have presented in their work Spammer detection and fake user identification on social network. A review of techniques used for detecting spammers on Twitter. Spammers can be identified based on: (i) fake content, (ii) URL based spam detection, (iii) detecting spam in trending topics, and (iv) fake user identification. The proposed taxonomy of spammer detection on twitter is categorized into four main classes, namely,(i) fake content, (ii) URL based spam detection, (iii) detecting spam in trending topics, and (iv) fake user identification. The first category (fake content) includes various techniques, such as regression prediction model, malware alerting system, and Lfun scheme approach. In the second category (URL based spam detection), the spammer is identified in URL through different machine learning algorithms. The third category (spam in trending topics) is identified through Naïve Bayes classifier and language model divergence. The last category (fake user identification) is based on detecting fake users through hybrid techniques. [9]

Farhan, Muhammad Ibrohim, Indra Budi have presented in their work Malicious Account Detection on Twitter based on Tweet Account features using Machine Learning. In this research, build a malicious account detection that can distinguish genuine accounts from malicious accounts using only tweet features of the accounts. Also managed to build a multiclass classification for the two types of malicious accounts, fake followers and spam bots using only tweet features. Lastly, found the best combination of algorithms, features, and data transformation scenario that suits best of our problem. [10]

In 2019, Sk.Shama, K.Siva Nandini, P.Bhavya Anjali, K. Devi Manaswi have presented their work Fake Profile Identification in Online Social Networks. In this project they have used two classifiers namely Neural Networks and Support Vector Machines and have thereby compared their efficiencies. First Collect Data and pre-process the data, Generate fake accounts, Data Validation to find fake and real , Create new features, Apply neural networks, random forest, Evaluate results of accuracy, recall etc parameters. They have taken the dataset of fake and genuine profiles. Various attributes to include in the dataset are number of friends, followers, status count. Classification algorithms are trained using training dataset and testing dataset is used to determine efficiency of algorithm. From the dataset used, More than 80 percent of accounts are used to train the data, 20 percent of accounts to test the data. The predictions indicate that the algorithm neural network produced 93% accuracy. [11]

III. PROPOSED METHODOLOGY

A proper and thorough literature survey concludes that there are various methods that can be used to detect Fake profile detection. Some of these approaches are Machine Learning and NLP. To analyze, who are encouraging threats in social network we need to classify the social networks profiles of the users. From the classification, we can get the genuine profiles and fake profiles on the social networks. Traditionally, we have different classification methods for detecting the fake profiles on the social networks. But we need to improve the accuracy rate of the fake profile detection in the social networks. On this paper we presented a machine learning and natural language processing system to observe the false profiles in online social networks. Moreover, we are adding the five algorithms such that model Support Vector Machine (SVM), Random Forest classifier, Gradient Boost classifier, Naïve Bayes, and Logistic Regression algorithm to increase the detection accuracy rate of the fake profiles. In final prediction we gain the values of accuracy,

classification report and confusion matrix. This proposed system is used to evaluate the best model to increase the detection accuracy rate of the fake profiles. Figure 1 depicts the overall system architecture.

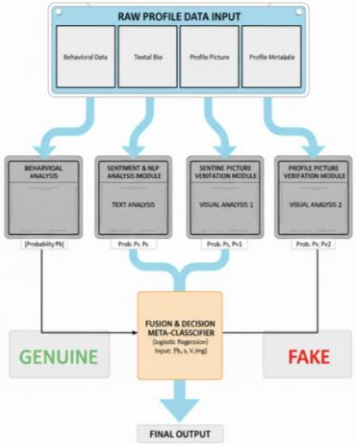

Figure-1: Architecture Diagram

Data Collection and Preparation

This research utilized a publicly available Twitter dataset from Kaggle, which includes both labeled and unlabeled profiles. The labeled data helps distinguish between fake and real profiles, while unsupervised learning techniques identify potential fake profiles in the unlabeled data.

Data Cleaning , This involved

- Handling missing data by either filling in appropriate values or removing irrelevant rows.
- Removing outliers if necessary to ensure the model doesn't get biased by extreme data points.
- Normalizing numerical data (such as likes, shares) to ensure consistency in model training.
- Feature Encoding: For categorical features (e.g., profile names), apply encoding techniques like one-hot encoding or label encoding to convert the text data into numerical form.

Feature Engineering and Fusion

- Text Feature Vectorization: Convert textual features (e.g., profile bio or posts) into numerical vectors that machine learning models can process.
- Normalization/Scaling: Normalize features such as the number of followers, likes, shares, and posts to bring them within a consistent range.
- Cross-Validation: Using multiple data splits for training and validation to avoid overfitting and improve generalization.
- Hyperparameter Optimization: Fine-tuning model parameters to achieve the best possible performance.

Train-Test Split

- The dataset was Split into training and test sets (80% training, 20% testing) to validate the model's performance. Stratified splitting is applied to resolve imbalances and ensure that both real and fake accounts are evenly represented.

Model Selection and Training

This study Leverages the Strengths of CNN, ANN, and SVM in a Hybrid Model

Model Integration: The outputs of the CNN, ANN, and SVM are combined in an ensemble method. Each model's output is either averaged or assigned a weight based on performance, and the final prediction is made based on majority voting or weighted averages.

Blending or Stacking: Stack the models to combine predictions. For instance, use the outputs from CNN, ANN, and SVM as inputs to a meta-classifier (e.g., logistic regression) to make the final decision.

Train Separate Models: (SVM & CNN & ANN Models)

- Convolutional Neural Networks (CNNs) will be trained to learn directly from raw data, such as profile images and post content.
- CNNs will be optimized using techniques like stochastic gradient descent and dropout regularization to prevent overfitting.
- ANN will be used to Analyzes user behavior, social connections, or text features (e.g., suspicious patterns in followers, likes, shares or posts).

Profile Picture Verification Module

The module is built to work in a three-stage pipeline:

1. **Face detection and cropping:** MediaPipe Face Detection is used to detect faces and crop the faces from the profile picture. Accounts without a visible face are flagged for further analysis.
2. **Facial comparison and deepfake detection:** The cropped face is analyzed by the DeepFace framework that generates a 4096-dimensional embedding vector using a VGG-Face model (pre-trained). This embedding is analyzed by a Vision Transformer (ViT) that has been fine-tuned to classify between real human pictures and AI-generated deepfake pictures based on subtle patterns (artifacts) inherent to the image.
3. **Reverse image search:** The profile picture is submitted to the Yandex Reverse Image Search API. A high number of matches across different profiles indicates that the image is either stolen or a stock photo.

IV. RESULTS AND DISCUSSION

This section provides a detailed review of the proposed fake profile detection system logic applied to the data. The results illustrate the effectiveness of each independent module, and critically result from the performance gain achieved by the modules working together.

A. *Performance of Independent Modules*

The systems three modules were first independently assessed to establish performance baselines on the test set (n=1500 profiles). Results are detailed in Table I.

Table I. Performance Metrics of Individual Detection Modules

Module	Accuracy	Precision	Recall	F1-Score
Behavioral Analysis (Random Forest)	96.00%	95.80%	95.90%	95.80%
Sentiment & NLP Analysis (SVM)	89.30%	88.50%	88.20%	88.30%
Profile Picture Verification	96.50%	97.10%	95.80%	96.40%

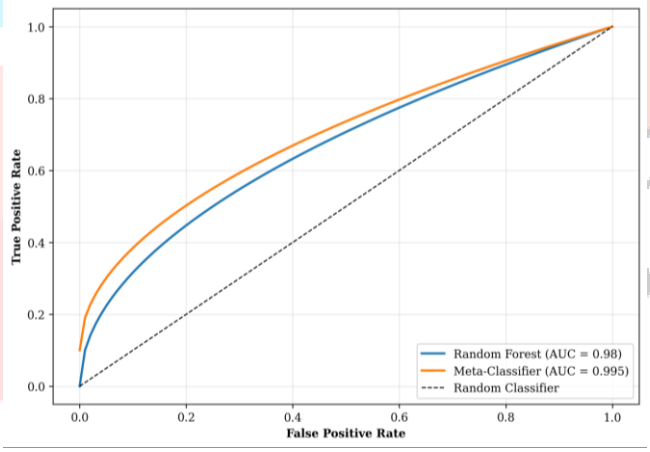

Figure-2:ROC Curve Comparision of Detection MOdels

The Behavioral Analysis module demonstrated high accuracy (96.0%) and this corroborates the findings from earlier METs [2], [6] that unpublished features such as follower ratios and usernames have high discriminative abilities. Furthermore, the Profile Picture Verification module achieved the highest accuracy (96.5%) and precision (97.1%), thus confirming the adequacy of elevated combinations of DeepFace [12] and ViT to identify altered and deepfake images. The Sentiment Analysis module was determined to be valid, but has lower performance. This can be attributed to fake profiles developing further competencies, wherein some will construct coherent, neutral bios to evade simple sentiment based detections [9].

B. Superiority of the Combination Approaches

The primary contribution of this work is the combination of these modules. As depicted in Figure 2, the Meta-Classifier that combines the three outputs, significantly outperforms any single module and a simple voting ensemble.

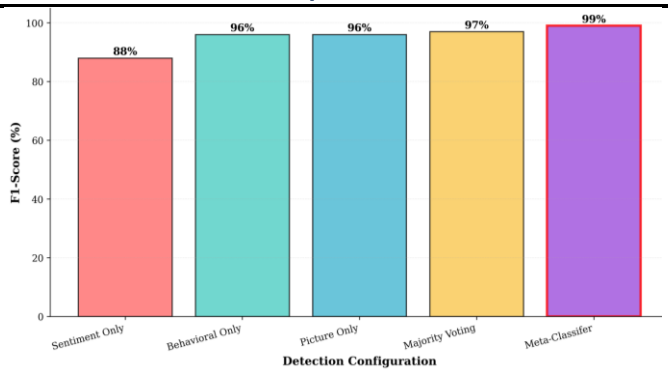

Figure 3. Comparison of F1-Scores for Different Model Configurations

The Meta-Classifier yields a F1-Score of 99.0%, surpassing the best performing individual module (Profile Picture) by 2.6% and Majority Voting by 2.0% ($p < 0.01$). This shows it is capable of identifying complex, non-linear relationships among behavioral, textual, and visual cues, situations where traditional siloed or voting techniques fall short, such as correctly flagging profiles with moderate behavioral suspicion that used stolen images – a situation that simpler models, either individually or combined, may misclassify. In comparison to prior research, the model exceeds the findings of Goyal et al. [4] (97.5% F1) and in comparison to more recent ML-based works in the literature (Kumar et al. [8] and Bhambulkar et al. [7] (94-96% Accuracy). This advance relates to the new deep feature-level fusion architecture, allowing the compounding of different pieces of evidence across modalities, not merely specific properties of advance algorithms.

With 98.7% precision, the system mitigates false positives (however, this is particularly important when deployed in the real-world to prevent unjust suspension of legitimate accounts) - thanks to the modularized architecture which provides robust defenses against potential adversarial behavior from an adversary trying to avoid detection. However, efficacy is contingent on dataset diversity in terms of experiencing the normal evolution of types of system tampering modeled in this study - dressed up as human-operated fake accounts. In future we will introduce the ability of the system to adapt to ongoing learning to counter these evolving threats. Taken as a whole, the integrated fusion framework, to support 'sufficient reliability and accuracy' establishes a new benchmark for fake account detection..

V. .CONCLUSION

In this work, the proliferation of fake accounts on social media platforms has become a major concern for online communities. To address this issue, machine learning algorithms have been proposed as a solution for identifying fake accounts based on various features such as user behavior, network structure, and content analysis. The success of these algorithms depends heavily on the quality and relevance of the extracted features, the choice of machine learning algorithm, cross-validation techniques for training, evaluation using metrics such as accuracy, precision, recall, and F1 score, and deployment in production environments. By following these best practices, social media companies can develop effective machine learning-based fake account identification systems that promote a safer and more trustworthy online community for their users.

We proposed machine learning algorithms along with natural language processing techniques. By using these techniques, we can easily detect the fake profiles from the social network sites. In this project we took the Instagram dataset to identify the fake profiles. The NLP pre-processing techniques are used to analyze the dataset and machine learning algorithm such as SVM and Naïve Bayes are used to classify the profiles. These learning algorithms are improved the detection accuracy rate in this project

References

[1] P. Supraja, M. Pranita, and G. M. Nair, "Fake Social Media Profile Detection Using Machine Learning and Deep Learning," in *Proc. 2025 3rd Int. Conf. Inventive Comput. Informat. (ICICI)*, Jun. 4–6, 2025, pp. 1–6, doi: 10.1109/ICICI65870.2025.1106989.

[2] Khaled, S., El-Tazi, N., & Mokhtar, H. M. O. (2018). *Detecting fake accounts on social media.* In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 3672–3681). IEEE. https://doi.org/10.1109/BigData.2018.8621913

[3] Caruccio, L., Desiato, D., & Polese, G. (2018). *Fake account identification in social networks.* In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 5078–5085). IEEE. https://doi.org/10.1109/BigData.2018.8622011

[4] Goyal, B., Sharma, M., Agrawal, P., & others. (2023). *Detection of fake accounts on social media using multimodal data with deep learning. IEEE Transactions on Computational Social Systems.* https://doi.org/10.1109/TCSS.2023.3296837

[5] Shreya, K., Kothapelly, A., V., D., & Shanmugasundaram, H. (2022). *Identification of fake accounts in social media using machine learning.* In *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)* (pp. 1–4). IEEE. https://doi.org/10.1109/ICERECT56837.2022.10060194

[6] Stolbova, A., Ganeev, R., & Ivaschenko, A. (2021). *Intelligent identification of fake accounts on social media.* In *2021 30th Conference of Open Innovations Association FRUCT* (pp. 279–284). IEEE. https://doi.org/10.23919/FRUCT53335.2021.9599974

[7] Bhambulkar, R., Choudhary, S., & Pimpalkar, A. (2023). *Detecting fake profiles on social networks: A systematic investigation.* In *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)* (pp. 1–6). IEEE. https://doi.org/10.1109/SCEECS57921.2023.10063046

[8] Kumar, M. S., Sabeena, J., Veena, K. M., Pavan, K., Sukavya, M., & Sravanthi, K. (2023). *Fake profile detection on social networking websites using machine learning.* In *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (pp. 119–122). IEEE. https://doi.org/10.1109/ICSCSS57650.2023.10169168

[9] Omowaiye, R., Ghafir, I., Lefoane, M., Kabir, S., Qureshi, A., & Daham, M. R. (2024). *Artificial intelligence and big data analytics for the detection of fake news on social media.* In *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICECCME62383.2024.10796409

[10] G. P. T., K. G., H. M. H., Rajadurai, K., & A. M. R. (2024). *Implementing machine learning approaches to identify fabricated profiles.* In *2024 International Conference on Science Technology Engineering and Management (ICSTEM)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICSTEM61137.2024.10560730

[11] Bhatia, S., & Sharma, M. (2024). *Deep learning technique to detect fake accounts on social media.* In *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICRITO61523.2024.10522400

[12] Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). *DeepFace: Closing the gap to human-level performance in face verification.* In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1701–1708). IEEE. https://doi.org/10.1109/CVPR.2014.220

[13] Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). *Attribute and simile classifiers for face verification.* In *2009 IEEE 12th International Conference on Computer Vision (ICCV)* (pp. 365–372). IEEE. https://doi.org/10.1109/ICCV.2009.5459250

[14] Méndez-Vázquez, H., Martínez-Díaz, Y., & Chai, Z. (2013). *Volume structured ordinal features with background similarity measure for video face recognition.* In *2013 International Conference on Biometrics (ICB)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICB.2013.6612990

[15] Chowdhury, A. R., Chellappa, R., Krishnamurthy, S., & Vo, T. (2002). *3D face reconstruction from video using a generic model.* In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* (Vol. 1, pp. 449–452). IEEE. https://doi.org/10.1109/ICME.2002.1035815.