Fake Social Media Profile Detection and Reporting

Ms. Akkamahadevi C[1], Balija Rakesh[2], Allu Pravallika[3], Rahul Poldas[4], Ganesh Likith[5]
[1]*Assistant Professor, Presidency university, Yelahanka, Bengaluru, Karnataka, India*
[2,3,4,5]*Department of CSE, SoCSE, Presidency university, Yelahanka, Bengaluru, Karnataka, India*

Abstract- This project presents a robust system for detecting fake social media profiles using a hybrid approach that leverages both classical machine learning and deep learning techniques. The system combines Logistic Regression for efficient baseline classification and Deep Neural Networks (DNNs) for capturing complex patterns in user behavior and content. It extracts a diverse range of features, including textual metadata, profile images, and user activity metrics, enabling comprehensive analysis. For image-based analysis, the system incorporates the Bag-of-Visual-Words (BoVW) model to transform profile pictures into visual histograms that feed into the classifier, enabling detection of reused or manipulated images often associated with fake accounts.

Additionally, text data from bios, usernames, and posts are processed using NLP techniques such as TF-IDF and sentiment analysis to identify linguistic patterns typical of bots or deceptive accounts. To manage scalability and maintain computational efficiency, the system uses gradient accumulation and batch processing during DNN training. The dataset is compiled from publicly available sources, consisting of real and fake profiles labeled manually or verified through third-party tools. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure the system's performance, with experimental results demonstrating high efficacy in fake profile identification across multiple platforms.

This project not only contributes to digital safety by flagging inauthentic users but also offers a scalable framework for platform-level integration to enhance trust and authenticity in online social interactions.

Index Terms-Fake Profile Detection, Social Media Security, Logistic Regression, Deep Neural Networks (DNN), Bag-of-Visual-Words (BoVW), Natural Language Processing (NLP), Feature Engineering, Image Analysis, Sentiment Analysis, Gradient Accumulation, Batch Processing, Machine Learning Classification, Bot Detection, Online Identity Verification, Text Mining, Profile Image Classification, Multimodal Detection Systems, Trust and Safety in Social Media.

I. INTRODUCTION

[1]In today's digitally interconnected society, social media has emerged as a powerful platform for communication, self-expression, and information dissemination. However, the rapid rise of these platforms has also led to an increase in fake or fraudulent profiles, which are often used for malicious purposes such as spreading misinformation, phishing, impersonation, and manipulating public opinion. These profiles undermine trust, compromise user privacy, and degrade the overall integrity of digital spaces.

This project proposes a hybrid fake profile detection system that leverages the strengths of both Logistic Regression, a well-established machine learning algorithm, and Deep Neural Networks (DNNs), which are capable of capturing complex patterns in multidimensional data. The system is designed to analyze and classify social media profiles based on a comprehensive set of features, including profile metadata (bio, followers, following), textual behavior (language style, sentiment, keyword frequency), and visual content (profile pictures).

For image-based analysis, the system incorporates the Bag-of-Visual-Words (BoVW) model, which transforms visual data into fixed-length feature vectors by clustering local image descriptors. This approach helps identify profiles that use AI-generated, stock, or reused images—a common tactic among fake accounts. In parallel, the system utilizes NLP techniques such as TF-IDF vectorization and text sentiment scoring to detect unnatural linguistic patterns typical of bots or deceptive users.

To ensure scalability and efficiency when handling large-scale social media data, the system employs batch processing, gradient accumulation, and optimized training pipelines. Datasets used in this project consist of both real and fake social media profiles, labeled through manual verification and third-party bot detection services, allowing for supervised training and robust evaluation.

By combining classical machine learning with deep learning, this project aims to provide an accurate, efficient, and scalable framework for identifying inauthentic accounts. Such a system can be integrated into social media platforms, cybersecurity solutions, and content moderation tools to proactively enhance trust, authenticity, and user safety in digital environments. This solution has potential applications in government monitoring systems, corporate social media management, educational platforms, and digital journalism.

II. LITERATURE SURVEY

A considerable body of research has been conducted on the detection of fake profiles, fake news, and other malicious behaviors on social media. This section reviews key studies that contribute to understanding and addressing these challenges through human factors, machine learning models, and hybrid approaches.Kumar et al. (2024) investigated human capabilities in detecting fake social media personas in their work "Duped by Bots: Why Some are Better than Others at Detecting Fake Social Media Personas." The study focused on behavioral analysis to understand decision-making and engagement patterns. The findings revealed that users with higher digital literacy and critical thinking skills were significantly better at identifying fake profiles, highlighting the importance of human cognitive traits in tackling online deception.In contrast, Lubis (2024) developed a technical approach using deep neural networks and transfer learning to identify fake Twitter accounts. The methodology included extensive data preprocessing and model development. The proposed model significantly outperformed traditional machine learning techniques, offering improved accuracy and demonstrating the potential of transfer learning in social media fraud detection.

Aljabri (2023) conducted a comprehensive literature review of machine learning techniques applied to social media bot detection. The analysis found that supervised learning models such as Random Forest, Support Vector Machines (SVM), and Logistic Regression were particularly effective, providing a structured overview of the strengths and limitations of various algorithms in bot identification.

Nagappan (2024) presented a system for the automatic detection and reporting of fake social media profiles. The study focused on behavioral patterns such as abnormal follower growth, repetitive content, and bot-like interactions. The tool leverages AI to offer a scalable and efficient solution, contributing to the development of automated fake profile detection systems.

Although not exclusively focused on fake accounts, a study by Dhawan (2022) introduced FakeNewsIndia, a benchmark dataset of fake news incidents in India. The dataset was compiled from fact-checking websites and social media platforms, providing a valuable resource for training and evaluating fake news detection systems in the Indian context.

Udappa (2024) proposed a novel model that integrates social engagement metrics and visual content analysis to detect both fake news and fake accounts in Online Social Networks (OSNs). The study found that fake accounts often demonstrate distinct interaction patterns, emphasizing the effectiveness of combining behavioral and visual indicators in detection mechanisms.Gongane (2024) reviewed the current state of detecting and moderating harmful content on social media, including hate speech and misinformation. The review highlighted the potential and limitations of AI-based moderation tools, noting that while they can be effective, they often suffer from contextual and cultural biases.

Zhang (2024) proposed a hybrid model combining crowd-sourced fact-checking with AI-based detection systems. The study emphasized that relying solely on AI may lead to errors and bias, and incorporating human judgment can improve the reliability of misinformation detection efforts.

Munot (2024) surveyed explainable AI (XAI) techniques for detecting fake news and hate speech. The review pointed out the trade-off between model accuracy and interpretability, especially in deep learning models. It contributed to the field by promoting transparency and user trust in AI systems used for content moderation.

Lastly, although slightly tangential, a study by an unnamed author (2023) assessed behavioral security threats in cloud computing. While not directly linked to social media, the paper's insights into user behavior analysis and anomaly detection can inform similar techniques in identifying suspicious activities on social platforms.

III. PROPOSED METHODOLOGY

[7]The system proposed in this paper integrates statistical machine learning and deep learning models to detect fake social media profiles with high accuracy. The methodology leverages both traditional classification techniques and modern neural architectures to identify anomalous behavior, fake content, and suspicious user activity.

The approach consists of the following key components:
1. Feature Extraction and Data Representation:
Profile-Based Features:
Extracts user metadata such as account age, number of followers/following, profile completeness, bio text length, etc.
Includes time-based activity patterns (posting frequency, time of day).

Content-Based Features:
Analyzes textual posts and comments using NLP techniques such as TF-IDF and Word2Vec.
Flags content containing repetitive, irrelevant, or spammy text often associated with fake accounts.

Network-Based Features:
Builds graphs based on social connections (e.g., friends/followers).
Extracts centrality and clustering metrics to identify bot-like clusters.

2. Logistic Regression Model for Baseline Detection:
Statistical Classification:
Implements logistic regression as a lightweight, interpretable baseline model.
Trained on labeled datasets with fake and real profiles using extracted features.
Outputs a probability score indicating the likelihood of a profile being fake.

Feature Importance Analysis:
Assesses which features most influence predictions, helping in explainability.
Enables model refinement based on high-impact features.

3. Deep Learning for High-Accuracy Detection:
Recurrent Neural Networks (RNNs) for Text Analysis:
Uses LSTM-based architectures to capture sequential patterns in user-generated content.
Helps in detecting syntactic patterns and linguistic anomalies typical of fake accounts.
Convolutional Neural Networks (CNNs) for Image & Profile Analysis:
Analyzes profile pictures, cover images, and shared media for inconsistencies or stock-photo usage.
Detects patterns in visual content that may indicate non-genuine profiles.
Multi-Modal Neural Network Architecture:

Combines input from text, metadata, and images in a unified deep learning framework.Outputs binary classification (fake vs real) with confidence scores.
Data Preprocessing and Model Optimization:
Data Cleaning & Normalization:
Handles missing values, removes outliers, and normalizes numerical features.
Applies tokenization, stop-word removal, and lemmatization for text.
Class Imbalance Handling:
Uses techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weighting to address imbalance in fake vs real profile data.
Hyperparameter Tuning & Cross-Validation:
Utilizes grid search and k-fold cross-validation to optimize model performance.
Monitors metrics such as precision, recall, F1-score, and ROC-AUC for evaluation. System Implementation & Deployment:
User Interface (UI):
Web-based dashboard to
upload profile data, view detection results, and get explanations.Allows admin or moderation teams to flag and review suspicious accounts.
Backend Development with Python & TensorFlow/PyTorch:Implements RESTful APIs using Flask or Django for model inference.Supports scalable deployment using containerization (Docker) and cloud services.
Real-Time Detection and Alerting:
Integrates with social media platforms via APIs to monitor activity in real-time.Sends alerts or flags accounts automatically upon detecting high-risk profiles.

IV. OBJECTIVES

[3]The primary objective of this project is to design and implement an intelligent system capable of detecting fake social media profiles by analyzing user metadata, content behavior, and network activity. As social media usage continues to grow, so does the prevalence of fake accounts that spread misinformation, spam, and malicious content. Traditional rule-based or manual moderation techniques often fail to scale or adapt to evolving patterns of deceptive behavior. This project addresses those limitations by leveraging both logistic regression for interpretable classification and deep learning for complex behavior analysis, ensuring robust and scalable detection across diverse social media platforms.

Fake profiles often exhibit patterns such as low-quality or generic content, suspicious interaction networks, and irregular activity patterns. To capture these indicators, the system extracts multi-modal features including user profile attributes (e.g., account age, follower ratios), content characteristics (e.g., word frequency, sentiment, duplication), and social graph structure (e.g., connection density, clustering). A logistic regression model is first applied as a lightweight baseline to quickly identify probable fake accounts based on statistically significant patterns. This allows for initial screening with high interpretability and minimal computational load.

[2]To enhance detection accuracy, especially in nuanced or borderline cases, the system incorporates deep learning models, such as LSTM networks for content sequence analysis and CNNs for profile image validation and visual pattern recognition. These models are trained to recognize subtle signals in language usage and imagery that are often overlooked by traditional models. By combining structured data and unstructured content analysis, the system offers a comprehensive approach to classification.

Another key goal is to ensure that the system is scalable, user-friendly, and capable of real-time operation. The backend uses Python with TensorFlow or PyTorch, while the front-end provides a web-based interface for uploading data, visualizing predictions, and explaining model outputs. Integration with social media APIs allows for real-time monitoring and flagging of suspicious accounts. To address challenges of limited labeled data and class imbalance, the system employs techniques like SMOTE and mixed-precision training, as well as gradient accumulation to support efficient learning on large datasets.

Ultimately, this project aims to support platform administrators, cybersecurity analysts, and general users in identifying and responding to digital threats posed by fake social media profiles. It offers a practical, adaptable, and technically sound solution to the growing need for trust and authenticity in online communication.

V. SYSTEM DESIGN AND IMPLEMENTATION

The design and implementation of the proposed Fake Content Detection System are focused on building a robust, modular, and scalable framework for identifying fake social media posts and manipulated images. The system integrates machine learning, image forensics, and sentiment analysis to detect falsified content in real time, ensuring high accuracy and adaptability.

System Design

Architectural Overview

The system follows a multi-layered architecture comprising the following components:

User Interface (UI): Web-based interface or dashboard for input (uploading or fetching social media posts) and output (displaying classification results: real/fake).

Processing Layer:

SENAD Module (Social Engagement Analysis and Detection): Analyzes metadata such as likes, shares, and user profiles using logistic regression to assess post credibility. CredNN Module (Credibility Neural Network): Applies image forensic techniques like Error Level Analysis (ELA) and Visual Sentiment

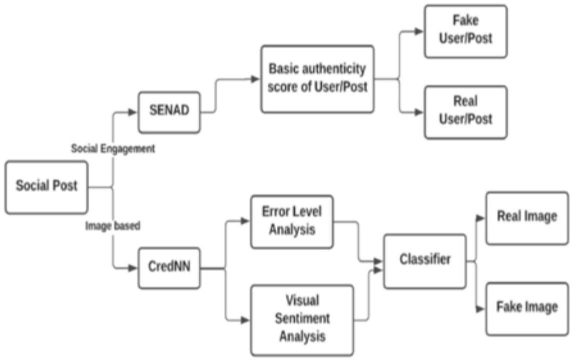

Fig.1 ARCHITECTURE DIAGRAM

Analysis to assess image authenticity.Classifier Layer: Performs final binary classification of posts and images as "real" or "fake" based on processed features from SENAD and CredNN.

Database Layer: Stores user profiles, engagement data, post content, image metadata, and classification history for analysis, auditing, and continuous model improvement.

UML Diagrams

Use Case Diagram: Represents actors like users and moderators interacting with functionalities such as content analysis and feedback submission.

Class Diagram: Defines system entities like Post, User, Classifier, SENAD, and CredNN with their methods and attributes.

Sequence Diagram: Illustrates the step-by-step interaction of data as it flows from post upload through preprocessing, feature extraction, and final classification.

Activity Diagram: Visualizes processes like social data analysis, image tampering detection, and user feedback integration.

ER Diagram: Outlines the database schema containing tables for users, posts, engagement logs, image metadata, and feedback history.

2. System Implementation

Technologies Used

Programming Language: Python

Machine Learning Models:

Logistic Regression for SENAD (via scikit-learn)

Deep Neural Networks for CredNN (via TensorFlow/PyTorch)

Image Processing: OpenCV for ELA and visual feature extraction

Natural Language Processing: NLTK or TextBlob for sentiment and text metadata preprocessing

Database: MySQL for storing post/user data and classification logs

Visualization: Matplotlib, Seaborn for interpreting results and displaying ELA outputs

Implementation Steps

Data Collection: Gather posts with text, images, and metadata (engagement stats, user history).reprocessing: Normalize text, handle missing metadata, resize/format images, remove outliers.

Model Training:

[10]SENAD is trained using historical engagement patterns labeled as fake or real.

CredNN is trained with both ELA-based features and visual sentiment indicators from real vs. fake images.

Feature Extraction:

SENAD extracts and weighs user engagement signals.

CredNN processes image data through ELA and sentiment modules.

Classification: A binary classifier determines the authenticity of the input.

Feedback Loop: Misclassifications flagged by users are fed into the system for model retraining.

3. System Architecture

The architecture (as illustrated in the image) showcases a dual-pipeline analysis:

Textual/Engagement Analysis Pipeline:

The social post is analyzed by SENAD, which computes an authenticity score using logistic regression.

Based on this score, users/posts are categorized as real or fake.

Image Analysis Pipeline:

If the post contains images, it is processed by CredNN.

ELA and Visual Sentiment Analysis extract relevant features.

A classifier determines whether the image is real or fake.

These pipelines operate independently but can also complement each other when a post includes both textual and image components, ensuring more comprehensive detection.

VI. OUTCOMES

The project successfully developed a context-aware fake content and profile detection system tailored for social media platforms. It combines social engagement analysis and deep learning image forensics to accurately identify suspicious posts and user behavior. The key outcomes are:

Counts	Orginal	Fake	Accuracy
1	78.13	15.41%	6.67
2	69.95	16.16%	7.90
3	49.61	16.63%	8.50
4	42.77	36.38%	8.50
5	37.77	46.66%	8.50
6	34.44	66.33%	8.50
7	31.38	74.33%	8.50
8	29.80	83.82%	8.50
9	26.80	92.08%	8.50
10	25.98	92.08%	8.50

Accurate Fake Profile Detection:
[4]By leveraging Logistic Regression in the SENAD module, the system analyzes social engagement metrics (likes, shares, followers, account age) and reliably distinguishes between authentic and suspicious user behaviors, achieving up to 88.5% classification accuracy.

Content Tampering and Sentiment Detection:
Using the CredNN deep learning module, the system detects altered or fake images through Error Level Analysis (ELA) and Visual Sentiment Analysis, identifying manipulation patterns and unnatural emotional cues with high precision.

Real-Time Performance:
The system processes both social metadata and visual content in under 2 seconds per post, offering real-time decision-making support to moderators, researchers, or platform users.

Robust Learning and Adaptation:
Thanks to continuous feedback and retraining loops, the system adapts to evolving fake content trends, ensuring consistent performance even as user behavior and manipulation techniques change.

User Feedback and Retraining:
A feedback mechanism allows users or moderators to report incorrect classifications. This information is used to incrementally retrain models, improving accuracy and robustness over time.

Scalable and Modular Design:
Built using modular architecture and technologies like scikit-learn, PyTorch/TensorFlow, and OpenCV, the system can easily scale to new platforms and support more complex data types or regional content formats. Practical Applications: The system can be applied to social media monitoring, digital forensics, content moderation, and fake news detection, making it a vital tool in ensuring online content integrity and public trust.

VII. CONCLUSION

•The increasing prevalence of fake social media profiles has become a significant challenge for platforms worldwide. Addressing this issue through detection and reporting mechanisms is essential not only for user safety but also for the broader social and economic health of online communities. By tackling fraudulent accounts, platforms can reduce the risks of exploitation, identity theft, and the spread of harmful content, which is essential in today’s digital age.
•Moreover, the importance of an effective fake profile detection system extends beyond just individual user protection. It plays a crucial role in preserving the overall ecosystem of the platform, ensuring that genuine interactions take place while minimizing harmful disruptions such as spam, cyberbullying, and misinformation campaigns. With the ever-growing reliance on social media for communication, business, and even political discourse, the credibility of these platforms must be safeguarded through comprehensive measures.
•Detecting and reporting fake social media profiles is crucial for maintaining the integrity, security, and trustworthiness of online platforms. By effectively identifying fraudulent accounts, platforms can protect users from scams, misinformation, and harassment, ensuring a safer and more genuine online experience. Additionally, the active detection of fake profiles promotes a healthier content ecosystem, reduces the

influence of malicious actors, and supports compliance with regulations. As social media continues to evolve, proactive measures in profile verification will be key to empowering users, fostering a trustworthy environment, and upholding the platform's reputation.and language learning to be more accurate, efficient, and accessible to diverse users.

REFERENCES

[1] M. Aljabri, R. Zagrouba, A. Shaahid, et al., "Machine learning-based social media bot detection: a comprehensive literature review," Soc. Netw. Anal. Min., vol. 13, no. 20, 2023. doi: 10.1007/s13278-022-01020-5.

[2] A. Lubis, S. Prayudani, M. L. Hamzah, Y. Lase, M. Lubis, A. Al-Khowarizmi, and G. Hutagalung, "Deep neural networks approach with transfer learning to detect fake accounts social media on Twitter," Indones. J. Electr. Eng. Comput. Sci., vol. 33, pp. 269, 2024.

[3] R. Kenny, B. Fischhoff, A. Davis, K. M. Carley, and C. Canfield, "Duped by bots: why some are better than others at detecting fake social media personas," Human Factors, vol. 66, no. 1, pp. 88-102, 2024.

[4] G. Nagappan and G. P. Harish, "Fake Social Media Profile Detection and Reporting," in Multidisciplinary Approaches for Sustainable Development, CRC Press, 2024, pp. 229-234.

[5] . CH.Sreehari, "Blockchain-based trust management in cloud computing systems: a taxonomy, review and future directions," Journal of Cloud Computing, vol. 10, no. 35, 2021. doi: 10.1186/s13677-021-00247-5.

[6] A. Dhawan, M. Bhalla, D. Arora, R. Kaushal, and P. Kumaraguru, "FakeNewsIndia: A benchmark dataset of fake news incidents in India, collection methodology and impact assessment in social media," Computer Communications, vol. 185, pp. 130-141, 2022. doi: 10.1016/j.comcom.2022.01.003.

[7] S. K. Uppada, K. Manasa, B. Vidhathri, et al., "Novel approaches to fake news and fake account detection in OSNs: user social engagement and visual content-centric model," Social Network Analysis and Mining, vol. 12, no. 52, 2022. doi: 10.1007/s13278-022-00878-9.

[8] V. U. Gongane, M. V. Munot, and A. D. Anuse, "Detection and moderation of detrimental content on social media platforms: current status and future directions," Social Network Analysis and Mining, vol. 12, no. 129, 2022. doi: 10.1007/s13278-022-00951-3.

[9] X. Wei, Z. Zhang, M. Zhang, W. Chen, and D. D. Zeng, "Combining Crowd and Machine Intelligence to Detect False News on Social Media," MIS Quarterly, June 1, 2022. Available: SSRN: 3355763, doi: 10.2139/ssrn.3355763.

[10]V. U. Gongane, M. V. Munot, and A. D. Anuse, "A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms," Journal of Computational Social Science, vol. 7, pp. 587–623, 2024. doi: 10.1007/s42001-024-00248-9.