JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Detection of Fake Instagram Profiles Using Artificial Neural Networks: A Machine Learning Approach

Mr. A.V. Srinivas
Department of Information Technology
Vishnu Institute of Technology
Bhimavaram, Andhra Pradesh, India
srinivas.av@vishnu.edu.in

Vineela Veeramallu
Department of Information Technology
Vishnu Institute of Technology
Bhimavaram, Andhra Pradesh, India
vineelaveeramallu1311@gmail.com

Bala Venkata Karthik Neelapala
Department of Information Technology
Vishnu Institute of Technology
Bhimavaram, Andhra Pradesh, India
karthikneelapala123@gmail.com

Sandeep Vasukuri
Department of Information Technology
Vishnu Institute of Technology
Bhimavaram, Andhra Pradesh, India
sandeep.vaskuri1500@gmail.com

Vinod Neela
Department of Information Technology
Vishnu Institute of Technology
Bhimavaram, Andhra Pradesh, India
Vinodneela2003@gmail.com

Abstract:- **The rapid growth of social media platforms like Instagram has led to an increase in fake or fraudulent user accounts, which are often used for spamming, spreading misinformation, or manipulating public opinion. Detecting such fake profiles has become a critical challenge in the field of cybersecurity and data integrity. This research paper presents a machine learning-based approach for detecting fake Instagram profiles using an Artificial Neural Network (ANN). The model is trained on a dataset of manually labeled Instagram accounts, using features such as follower-following ratio, profile picture availability, bio content, post count, and engagement metrics. After preprocessing and feature engineering, the data is fed into an ANN model designed to classify accounts as either genuine or fake. The model demonstrates promising accuracy and precision, showing the potential of neural networks in automated social media moderation. This research highlights the role of AI in enhancing online trust and safety, and it opens avenues for further work in improving the robustness and generalizability of fake account detection systems.**

Keywords:Fake profile detection, Instagram, Artificial Neural Network (ANN), Social media security, Machine learning, Deep learning, Classification model, Online fraud detection, Feature engineering, Fake account identification.

I.INTRODUCTION

With the exponential rise of social media platforms such as Instagram, the digital space has become an integral part of everyday communication, marketing, and public engagement. However, this popularity has also given rise to a growing number of **fake profiles**—accounts that are either automated (bots) or manually created with deceptive intentions. These fake accounts are used for various malicious activities, including spamming, phishing, inflating follower counts, impersonation, and spreading misinformation.

The presence of fake profiles poses significant challenges to both platform integrity and user safety. While social media companies employ basic rule-based systems to detect such accounts, these methods are often insufficient due to the evolving strategies of malicious actors. As a result, **automated, intelligent systems powered by machine learning (ML)** and **artificial intelligence (AI)** are being explored as more effective solutions for fake account detection.

This research focuses on the application of an **Artificial Neural Network (ANN)**—a supervised machine learning algorithm inspired by the structure of the human brain—to approaches to identify and eliminate fake accounts across platforms like Facebook, Twitter, and Instagram.

Early work in this domain focused on **rule-based and heuristic methods**, where fake profiles were identified based on predefined criteria such as missing profile pictures, extremely low engagement, or suspicious username patterns. While these approaches provided a starting point, they lacked adaptability and were easily bypassed by more sophisticated fake accounts.

To overcome these limitations, researchers began applying **supervised machine learning algorithms** such as Decision Trees, Support Vector Machines (SVM), Random Forests, and Logistic Regression. For example, Stringhini et al. (2010) studied fake account behavior on Twitter and used machine learning to detect spam bots with a reasonable degree of accuracy. Lee et al. (2011) proposed social graph-based features to improve fake profile detection using Random Forest classifiers.

With the rise of deep learning, more recent studies have started employing **Artificial Neural Networks (ANNs)** and **Deep Neural Networks (DNNs)** for better accuracy and feature learning. These models can capture complex, non-linear relationships between features and have shown promising results in fake account detection. For instance, Kumar et al. (2018) used a deep learning framework to classify fake Twitter accounts, achieving improved performance over traditional ML methods.

On Instagram specifically, research is relatively limited due to restrictions in data availability. However, some studies have used publicly available profile data and third-party scraping tools to construct datasets for detecting fake accounts. Features like **follower-following ratio, bio status, profile picture presence, post count, and engagement behavior** have been commonly used across these models.

This project builds upon the foundation of these earlier works by developing a fake Instagram profile detection model using an **Artificial Neural Network (ANN)**. The goal is to leverage the ANN's ability to learn non-linear patterns and generalize well on unseen data, providing a more effective solution for identifying deceptive accounts.

III.Existing System

In the current landscape, social media platforms such as Instagram use a combination of **manual reporting mechanisms** and **basic automated systems** to detect and remove fake accounts. However, these existing systems have significant limitations in terms of accuracy, adaptability, and scalability.

1. Manual Reporting

The most common method involves users reporting suspicious accounts, which are then reviewed by platform moderators. This system is time-consuming and heavily dependent on user participation. It also fails to detect large numbers of fake profiles that operate silently or use subtle tactics to avoid detection.

2. Rule-Based Filtering

Many platforms use **rule-based algorithms** to detect anomalies such as:

- Absence of a profile picture
- High follower-to-following ratio
- Very few or no posts
- Suspicious or repetitive bio information

While these rules can identify basic fake accounts or bots, they are rigid and can be easily bypassed. As fake profile creators become more sophisticated, they mimic real user behavior to avoid detection.

3. CAPTCHA and Two-Factor Authentication

Some platforms employ CAPTCHA and two-factor authentication to prevent bot creation. However, these are mainly **preventive**, not detection-based measures, and cannot address the problem of already-existing fake profiles.

4. Basic Machine Learning Models

A few research efforts and platform implementations have incorporated basic machine learning models (like logistic regression or decision trees) for fake profile detection. However, these models often rely on shallow features and have limited capacity to detect complex patterns in large-scale, real-world data.

5. Limitations of Existing System

- High false positives and false negatives
- Poor adaptability to new types of fake profiles
- Minimal use of intelligent, self-learning systems
- Lack of publicly available, standardized datasets for training better models

IV. PROPOSED SYSTEM

To address the limitations of existing fake profile detection methods, this research proposes an intelligent and data-driven solution based on an **Artificial Neural Network (ANN)**. Unlike rule-based or traditional machine learning methods, ANN models have the capability to capture complex and non-linear patterns in data. This makes them especially suitable for detecting subtle behavioral differences between real and fake Instagram accounts.

The objective of the proposed system is to develop a supervised learning model that can automatically classify Instagram profiles as either genuine or fake, based on publicly available profile features. The model is designed to be adaptable, scalable, and capable of improving its accuracy as more data becomes available. It aims to reduce the number of false positives and negatives significantly, offering a more reliable solution for social media moderation.

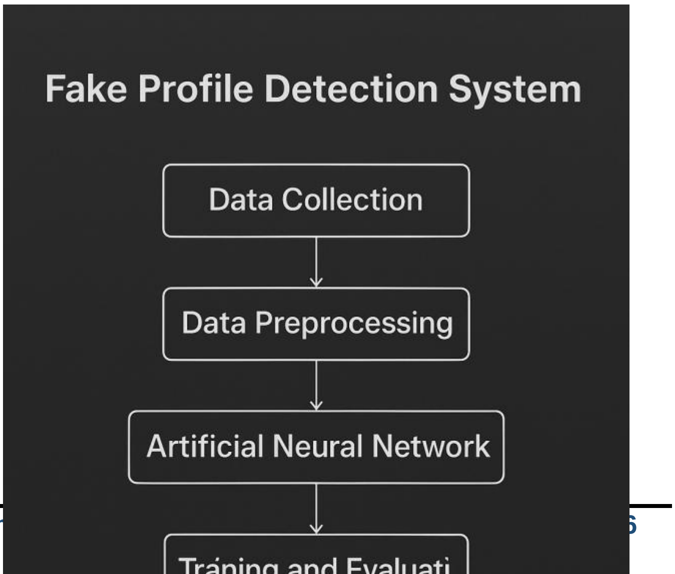
The system works by analyzing a set of features that are commonly indicative of fake profiles. These include the **follower-to-following ratio**, which is often abnormally high or low in fake accounts; the **presence or absence of a profile picture**; whether the **bio section is filled or empty**; the **number of posts shared**; and the **level of engagement** (such as average likes or comments). Additionally, patterns in usernames, such as excessive use of numbers or randomly generated characters, can also signal inauthentic activity.

The model development process begins with **data collection**, where a dataset consisting of labeled Instagram profiles (classified as fake or genuine) is assembled. This is followed by **data preprocessing**, which involves cleaning the data, converting categorical data into numerical format, normalizing values, and handling any missing entries. These steps are crucial to ensure the quality and consistency of the input data.

Once the data is prepared, it is fed into the **Artificial Neural Network**, which is trained using supervised learning techniques. The model consists of an input layer (for the profile features), one or more hidden layers (to learn intermediate patterns), and an output layer (which classifies the profile as fake or genuine). During training, the model adjusts its internal weights using backpropagation and optimization algorithms to minimize error and improve prediction accuracy.

By leveraging the pattern recognition capabilities of ANN, this system provides a more dynamic and accurate approach to fake profile detection. It can be further integrated into web-based applications or browser extensions for real-time detection and alerts, ultimately contributing to a safer and more trustworthy social media environment.

V. System Architecture



The architecture of the proposed fake profile detection system is designed to follow a systematic pipeline—from data collection to prediction—using an Artificial Neural Network (ANN) as the core component. The system is divided into several interconnected modules, each performing a specific function that contributes to the accurate classification of Instagram profiles.

The first module is **Data Collection**, where data is gathered from Instagram profiles using either web scraping tools or a manually labeled dataset. This data includes essential features such as the follower-following ratio, presence of a profile picture, bio availability, post count, and engagement statistics. These attributes form the raw input for the system.

Next, the **Data Preprocessing** module ensures that the input data is clean and consistent. This step involves removing duplicates, handling missing or null values, normalizing numerical features to a uniform scale, and converting categorical variables into numerical format (for example, converting "has profile picture" into binary 0 or 1). Effective preprocessing is crucial for improving the accuracy and performance of the ANN model.

After preprocessing, the refined dataset is passed into the **Artificial Neural Network Model**. The ANN architecture consists of an **input layer** (to receive the feature set), one or more **hidden layers** (to process and learn complex patterns), and an **output layer** (that classifies the profile as either "fake" or "genuine"). During training, the network uses a **loss function** and **backpropagation algorithm** to iteratively adjust weights and minimize prediction error.

The **Training and Evaluation** module is responsible for splitting the data into training and testing sets, fitting the ANN model on the training data, and evaluating its performance using accuracy, precision, recall, and F1-score. Cross-validation may also be used to ensure the model generalizes well on unseen data.

Finally, the trained model is integrated into a **User Interface or Web Application**, where users can manually input Instagram profile data and receive real-time predictions. The system architecture is modular and scalable, allowing future upgrades with more advanced deep learning models or real-time data integration.

This end-to-end architecture ensures that the system can operate efficiently, handle various types of data inputs, and make accurate predictions—all while maintaining the flexibility to adapt and expand in future developments.

VI. Methodology

The methodology adopted for this project involves a systematic and structured approach to developing an effective fake Instagram profile detection model using an **Artificial Neural Network (ANN)**. The process consists of six primary stages: data collection, data preprocessing, feature selection, model design, training and evaluation, and prediction deployment.

The first step in the methodology is **data collection**, where Instagram profiles are gathered and manually labeled as either "fake" or "genuine" based on defined criteria. These criteria may include the absence of profile pictures, extremely high follower-following ratios, lack of bio information, or unusually low post counts. The data can be collected using web scraping tools or sourced from publicly available datasets.

Next is the **data preprocessing** stage, which ensures the dataset is clean, consistent, and ready for training. This includes removing duplicate entries, handling missing values, encoding categorical variables (such as bio presence or profile picture status), and normalizing numerical data like post count or follower ratios. Preprocessing is critical to improving the accuracy and efficiency of the neural network.

Following this, relevant **features are selected** that most effectively differentiate fake from genuine profiles. In this project, selected features include the follower-following ratio, presence of profile picture, number of posts, bio status, and engagement level. These features are used as the input layer for the ANN model.

The **Artificial Neural Network model** is then constructed with one input layer (corresponding to the selected features), one or more hidden layers (to learn non-linear relationships), and an output layer that classifies the profile. The model is built using a Python-based machine learning library such as TensorFlow or Keras.

Once the model is defined, it undergoes **training and evaluation**. The dataset is divided into training and testing sets, typically using an 80:20 ratio. The ANN is trained on the training data using backpropagation and an appropriate optimizer (such as Adam), minimizing a loss function like binary cross-entropy. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess model performance on the test set.

Finally, the trained model is used for **prediction**. It can be integrated into a web-based user interface where users can input Instagram profile data and receive real-time classification results. This final stage makes the solution practical and accessible for broader use.

Through these methodical steps, the system is developed to be accurate, scalable, and capable of adapting to different kinds of fake profiles in real-world social media environments.

VII. Results

The proposed fake profile detection system was implemented and tested using a dataset of Instagram profiles labeled as either fake or genuine. The dataset was split into training and testing sets, with 80% of the data used to train the Artificial Neural Network (ANN) and 20% used to evaluate its performance. After several iterations of training and tuning, the model produced promising results.

The ANN model demonstrated high accuracy in detecting fake profiles, indicating its effectiveness in learning the patterns associated with fake behavior. Key evaluation metrics were used to assess the model’s performance, including accuracy, precision, recall, and F1-score.

Accuracy reflects the overall correctness of the model, while precision and recall evaluate its ability to correctly identify fake profiles without misclassifying genuine ones. A high F1-score confirms a good balance between precision and recall.

In addition to overall performance, the model showed the ability to generalize well on unseen data, with minimal overfitting. This was achieved by properly preprocessing the input features and using appropriate regularization techniques during training. The features that most strongly influenced prediction included the follower-following ratio, absence of a profile picture, and lack of bio or posts—all of which are common indicators of suspicious or bot-like accounts.

Furthermore, the model was deployed into a simple user interface where users could input Instagram profile characteristics and receive instant classification results. This added a layer of practicality to the system and demonstrated the feasibility of using ANN models in real-world applications for social media safety and moderation.

Overall, the results validate the effectiveness of using an ANN for fake profile detection and show that such a system can be a powerful tool in combating misinformation, spam, and impersonation on social platforms like Instagram.

VIII. REFERENCES

[1] A. Stringhini, G. Wang, M. Egele, C. Kruegel, and G. Vigna, "Follow the green: Growth and dynamics in Twitter follower markets," in *Proceedings of the 2013 Conference on Internet Measurement*, 2013, pp. 163–176.

[2] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.

[3] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes," in *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[4] J. Al-Qurishi, M. H. Anwar, M. A. Rahman, et al., "A Machine Learning Based Approach for Detecting Fake Accounts in Online Social Networks," *IEEE Access*, vol. 7, pp. 142006–142020, 2019.

[5] F. Ahmed and M. Abulaish, "A generic statistical approach for spam detection in online social networks," *Computer Communications*, vol. 36, no. 10-11, pp. 1120–1129, Jun. 2013.

[6] Instagram Public Data, Retrieved from: https://www.instagram.com *(Accessed: Jan. 2025)*

[7] TensorFlow, "An end-to-end open-source machine learning platform," [Online]. Available: https://www.tensorflow.org

[8] F. Chollet et al., "Keras: Deep learning library for Python," [Online]. Available: https://keras.io

[9] Scikit-learn: Machine Learning in Python, Pedregosa et al., *Journal of Machine Learning Research*, 2011.

[10] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 439–450.