Fake Instagram Profile Detection

Apoorva Neha
Dept. of Mathematics
230123008

Samriddhi Varshney
Dept. of Mathematics
230123057

Abstract

In the digital age, Online Social Networks like Instagram have become central to global communication, but they also face rising threats from fake profiles. These accounts often engage in harmful activities such as identity theft, cyberbullying, and misinformation. To address this, our project proposes a machine learning-based solution, leveraging Artificial Neural Networks (ANNs) and BERT for NLP, to identify fake profiles. By analyzing user behavior, activity patterns, and bio information, the system effectively distinguishes genuine users from fraudulent ones, contributing to safer and more trustworthy social media interactions.
Keywords: Online Social Networks, machine learning, artificial neural networks, NLP, social media security.

1. Introduction

Social networking sites have become integral to modern life, but the rise of fake accounts poses significant challenges. To address this, our project combines numerical features (like follower count and post frequency) and textual data (like bio content) using Artificial Neural Networks (ANN) and BERT. By analyzing both types of data, we can accurately detect fake profiles. The model is trained on a dataset of known fake and real profiles.

2. Methodology to Identify Fake Accounts

The process for detecting fake accounts involves several key steps, from data collection to model ensemble for final prediction:

- **Data Collection:** The first step is to scrape data from Instagram for real accounts, while fake account data is synthetically generated.
- **Feature Extraction:** Various features are extracted from the profiles, including numerical data such as follower count,num of posts, as well as textual features like bio content, profile description, and username structure.
- **ANN for Numerical Data:** A simple Artificial Neural Network (ANN) is applied to the numerical data to predict whether an account is real or fake.
- **NLP Models for Textual Data:** Various NLP methods, including BERT, are tested on textual features like bios and descriptions to classify accounts, with models evaluated for their accuracy in detecting fake profiles.
- **Model Ensemble:** After evaluating individual ANN and NLP models, their outputs are combined using an ensemble method. The ensemble model integrates the strengths of both ANN and NLP models to improve overall prediction accuracy.

3. Detection Strategy

The collected data is used to extract the following parameters for fake account detection:

1. Number/Length of username: A valid username is usually close to the user's real name and contains minimal numerical values.
2. Number/Length of full name: Fake accounts often lack a proper name.
3. Length of account description: Real accounts typically have a well-defined bio, whereas fake accounts often have minimal or no bio.
4. If the account is private or not: Private accounts tend to be more authentic, while fake accounts may opt for public visibility.

5. Total number of posts: Genuine accounts usually have a reasonable number of posts, while fake accounts may have fewer.
6. Total number of followers: Fake accounts often exhibit a disproportionate follower-to-post ratio.
7. Total number of following: Fake accounts may follow many others in an attempt to seem active.
8. Account Descriptio: Fake accounts often contain external URLs for promotional purposes.

4. Module Description

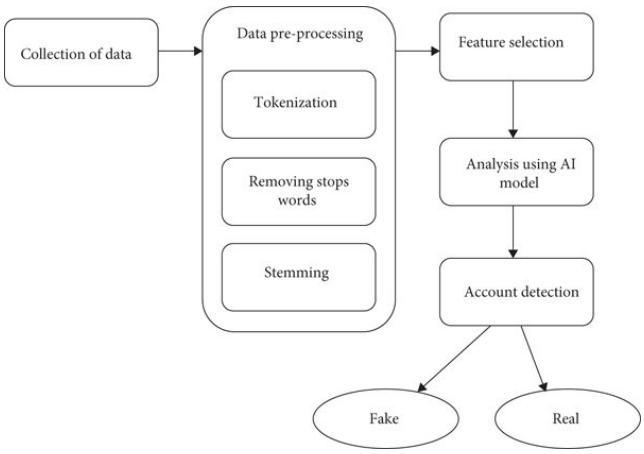
- **NumPy and Pandas: NumPy** provides fast operations for handling large arrays and matrices. **Pandas** offers powerful data structures for data cleaning, analysis, and manipulation.
- **TensorFlow and TensorFlow Hub: TensorFlow** is a deep learning framework used for building and training ML models. **TensorFlow Hub** provides pre-trained models like BERT for easy integration and reuse.
- **Matplotlib and Seaborn: Matplotlib** is a plotting library used for basic visualizations like line and bar charts. **Seaborn** builds on Matplotlib and simplifies the creation of statistical visualizations.
- **Scikit-learn (sklearn): Scikit-learn** provides simple and efficient tools for machine learning and data mining. It includes algorithms for classification, regression, and clustering.
- **Instaloader: Instaloader** is a Python tool to download Instagram data like bios, followers, and posts. It is used to scrape real profile data for building the training dataset.

5. Technologies Used

- **TF-IDF + Logistic Regression:** A fast and simple baseline for textual feature classification. It converts bios and descriptions into numerical form and uses logistic regression to classify fake vs. real accounts.
- **BERT Embeddings + Random Forest:** BERT generates deep contextual embeddings from textual data such as bios and descriptions. These embeddings are then fed into a Random Forest classifier, which makes robust predictions by aggregating the outputs of multiple decision trees.
- **Fine-tuned BERT:** This high-performing model is trained end-to-end on the fake account dataset. While powerful, it requires close monitoring to avoid overfitting and high computational costs.
- **Artificial Neural Networks (ANN):** ANN is used for numerical features like follower count, post frequency, etc. It captures complex non-linear patterns using layers of neurons with activation functions and dropout for generalization.
- **Logistic Regression:** Used as a meta-classifier in the ensemble model to combine predictions from multiple classifiers. It helps improve overall accuracy by learning how to best weight the outputs of base models.

6. Methodology

The methodology follows a structured pipeline to detect fake Instagram accounts by integrating both numerical and textual features:



1. **Data Collection and Exploration:** Real Instagram data is scraped using Instaloader, and fake profiles are synthetically generated. Data is explored using plots like histograms and heatmaps to identify trends and anomalies.
2. **Feature Extraction and Preprocessing:** Key numerical and textual features are ex-

tracted, including username structure, bio length, and follower counts. Text data is cleaned, and numerical features are standardized. Missing values are handled, and labels are encoded for training.

3. **Model Building – ANN:** A simple Artificial Neural Network is built using TensorFlow/Keras to classify accounts based on numerical features. The model includes dense layers, ReLU activations, and dropout layers to reduce overfitting.
4. **NLP-Based Text Classification:** Textual features (bios, names) are analyzed using models like TF-IDF with Logistic Regression, and BERT embeddings with Random Forest. Multiple models are evaluated to find the most accurate for textual classification.
5. **Ensemble Learning:** Predictions from the ANN (numerical) and NLP (textual) models are combined using Logistic Regression as a meta-classifier. This ensemble improves overall prediction accuracy by leveraging strengths of each model.
6. **Performance Evaluation:** Models are evaluated using metrics such as accuracy, precision, recall, and confusion matrix. Visualization of training loss, ROC curves, and prediction performance help assess model effectiveness.

7. Advantages of Proposed Work

The proposed work offers several advantages for fake Instagram profile detection. By using advanced deep learning models like BERT and ANN, the system effectively analyzes both numerical and textual features to improve accuracy. It also leverages ensemble learning and sentiment analysis to enhance robustness. This approach strengthens the detection of fake accounts, ensuring better social media security.

8. Future Directions

To enhance fake account detection, future work could explore:

- **Real-Time Detection:** Develop systems for quick identification and mitigation of fake accounts with low-latency inference.
- **User Education:** Educate users on fake account signs and improve reporting for better model training.
- **Interdisciplinary Approaches:** Combine insights from cybersecurity, psychology, and sociology for more effective detection.

9. Metrics of the model

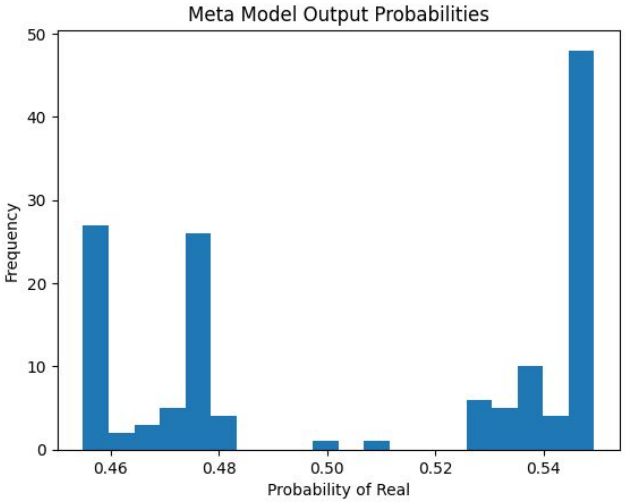


Table 1: Ensemble Model Evaluation and Meta-Model Weights

Metric	Value
Accuracy	0.9648
Precision	0.9467
Recall	0.9861
F1 Score	0.9660
ROC AUC Score	0.9958
Numerical Model Weight (w_1)	0.0881
NLP Model Weight (w_2)	0.2967
Bias (Intercept)	-0.1862
Numerical Model Trust	22.9%
NLP Model Trust	77.1%

References

1. ”Unmasking Deceit - A Fake Profile Detection using Artificial Neural Network” -Mr.Anbazhagan.R, Mr. Jayakrishna.B, Mr.Arun Kumar.S
2. Fake Instagram Profile Detection Model (Kaggle)- Durgesh Rao