
FAKE INSTAGRAM PROFILE IDENTIFICATION USING ML

Aniruddha Lalde^{*1}, Apurv Badave^{*2}, Nikhil Elajale^{*3}, Pankaj Godara^{*4},

Prof. Priyanka Kinage^{*5}

^{*1,2,3,4,5}Department Of Computer Engineering Smt. Kashibai Navale College Of Engineering
Pune, India.

ABSTRACT

The pervasive rise of social media has become an undeniable aspect of modern life, with its global dominance steadily increasing. However, this growth has also brought about a host of ecosystem challenges, including the proliferation of hate speech, fraudulent activities, and the spread of fake news. These issues, compounded by the staggering number of over 1.7 billion fake accounts across social media platforms, have already resulted in significant losses, and the process of removing these accounts remains daunting and time-consuming. With Instagram's user base expanding rapidly, there's a growing urgency to address the issue of identifying fake accounts on this platform. Traditionally, manual identification processes are laborious and time-intensive.

Keywords: Fake Profile Identification, User Authentication, Data Preprocessing, Model Training, Online Security, Machine Learning.

I. INTRODUCTION

Online Social Networks (OSNs) such as Facebook and Instagram have become increasingly pervasive in modern society, playing a vital role in communication and serving as platforms for personal and business promotion. Initially, an account's popularity is often gauged through metrics like follower count and engagement indicators such as likes, comments, and views on shared content. Consequently, users may resort to artificial means to inflate these metrics for personal gain. Common methods include the use of bots, purchasing social metrics like likes and followers, and utilizing platforms for metric trading. Notably, a significant number of Instagram accounts are automated, and bots have been known to generate more internet traffic than humans. Leveraging a comprehensive dataset comprising various features extracted from genuine and fraudulent accounts for employing advanced algorithms and feature engineering methodologies, we endeavor to construct a robust framework for automated detection, thereby enhancing the integrity and reliability of the Instagram ecosystem. Through rigorous experimentation and evaluation, our findings underscore the efficacy of the proposed system in differentiating between genuine and fake profiles, offering a promising solution to mitigate the growing menace of online deception.

II. DATASET & INPUT

The detection of fake accounts on social media platforms is crucial due to their role in artificially inflating the popularity metrics of other users. These accounts often exhibit characteristics such as a high number of accounts followed coupled with a low number of followers, erratic liking behavior, and identifiable traits like the absence of a profile picture and unusual usernames.

A. Features Of Dataset

To facilitate the detection of fake accounts, a dataset comprising of numerous real accounts & fake accounts was meticulously curated through manual labeling. Various factors were considered during data collection, including follower and following counts, media posting frequency, comments on media, and profile characteristics such as the presence of a profile picture and the format of the username.

In the dataset, the selected base features can be listed as below:

- Total media number of the account.
- Follower count of the account.
- Following count of the account.
- Number of digits present in account username.
- Whether account is private, or not (binary feature).

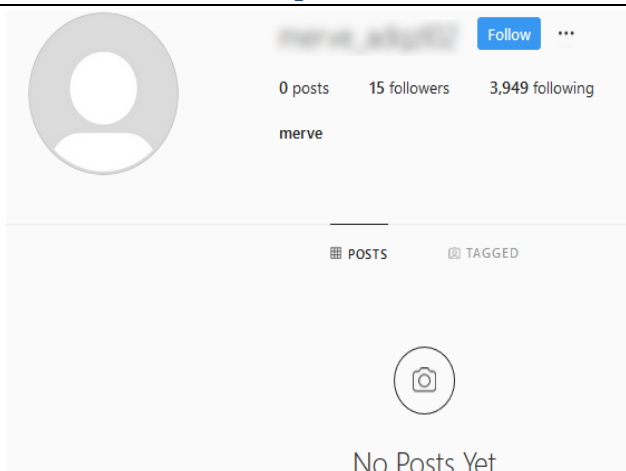


Fig 1: Fake account example from dataset.

B. Oversampling

Distribution of classes in the fake account dataset is not even. This results in poor performance for the outnumbered class. SMOTE oversampling technique is utilized to increase number of samples for fake accounts. K is chosen as 5 for this work. In the implementation of SMOTE, SMOTE-NC is applied which considers not only the quantity classes but also the categorical classes. After applying oversampling, all classifiers are trained on equal number of training samples per class.

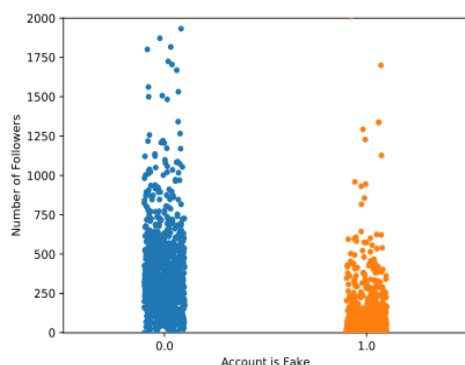


Fig 2: In-class data distributions for "following count" feature.

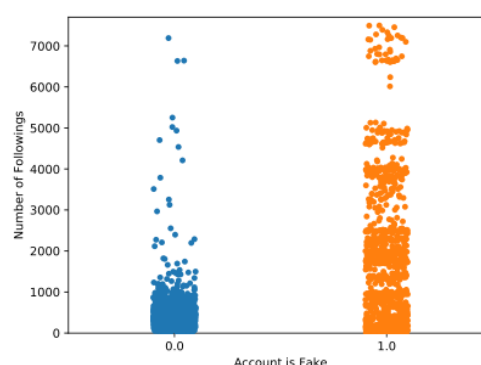


Fig 3: In-class data distributions for "follower count" feature.

III. FAKE ACCOUNT DETECTION

This section delves into the identification of automated accounts, commonly referred to as bots, which engage in automated activities such as following, liking, and commenting on content. These actions are typically targeted towards specific hashtags, locations, or followers of particular accounts with the aim of boosting their own popularity metrics. Automated accounts may exhibit behavior that is entirely inorganic or a combination of organic and inorganic actions. It's worth noting that the presence of organic behavior in such accounts can occur when users continue to pursue their own interests while the bot operates in the background.

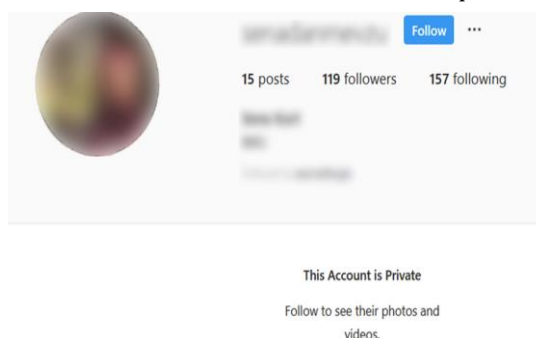


Fig 4: Example private account preview. Media details are not visible for private accounts.

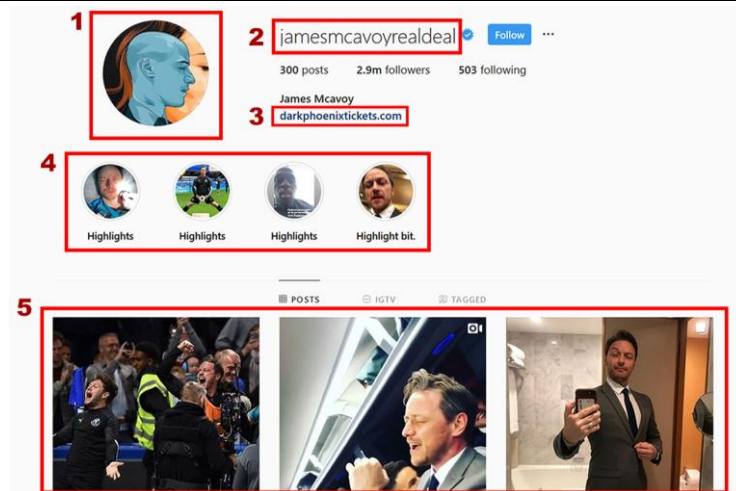


Fig 5: General preview of an Instagram profile.

- 1: Profile picture
- 2: Username
- 3: External URL
- 4: Highlight reel
- 5: User media.

A. Dataset Features & Methodology

To compile the authentic accounts, we selected individuals from our personal networks, primarily from our circle of friends. In contrast, to gather the automated accounts, we analyzed the source codes of the most widely used open-source Instagram bots, which are instrumental in generating artificial engagement. We identified specific behaviors characteristic of these bots to accurately label the accounts.

Hashtags serve as a prominent avenue for identifying inorganic activity. For instance, a common method for detecting fake likes on Instagram involves scrutinizing the usage of hashtags associated with like and follow trading. In our approach, we concentrated on targeting the most popular hashtags, as it was observed that this method facilitated the detection of automated behavior more efficiently and rapidly. If a user is found to engage in activities such as following and unfollowing within predefined time intervals, as determined by parameters from online Instagram automation tools, they are categorized as automated accounts. We utilized the Instagram API alongside a Python wrapper to meticulously gather comprehensive media and user information from the accounts over a six-month period. To safeguard user privacy, any personally identifiable information such as usernames, photos, comments, and hashtag details were deliberately excluded from the dataset.

Below are the scrapped base features from these accounts:

- Total media number of the accounts.
- Follower count of the account.
- Following count of the account.
- Number of photos user is tagged by someone else.
- Average recent media hashtag number.

If the account has no media, all features scrapped from user posts are assigned as 0. Additionally helpful features are derived using the base features such as:

- Average recent media like to comment ratio (LCR).
- Follower to following ratio (FFR).
- Whether account has not any media, or not.

B. Bias Problem

There existed some unfavorable bias within certain features of the dataset. Figures 6-7 display the distribution of the entire dataset concerning selected continuous features, where a label of "Fake Engagement" equal to 1

denotes accounts involved in fake engagement or automated behavior, while 0 corresponds to accounts with only genuine engagements or real accounts. These figures reveal bias within the dataset across the chosen features. While the bias in follower and following numbers appears unrealistic (likely due to unintentionally categorizing accounts with low follower and following counts as real accounts, which does not accurately reflect real-world scenarios), the bias in average hashtag usage per post appears more natural. Accounts engaged in automated behavior tend to employ more hashtags per post, contributing to this observed bias.

The distribution of the dataset is projected across selected binary features. These tables highlight the presence of bias within these features as well, although this time, the biases appear to be more in line with reality. For instance, the presence of highlight reels can be seen as an effective differentiator, as real engagement accounts predominantly lack this feature. Similarly, the absence of a URL in the profile can be considered a genuine bias, as it aligns with typical user behavior.

C. Cost Sensitive Feature Selection

To address these unrealistic biases and identify the most impactful features, a novel approach employing a cost-sensitive genetic feature selection algorithm has been devised. The algorithm, outlined in Algorithm 1, begins by normalizing the continuous features while leaving the binary features unchanged. Subsequently, the normalized data is inputted into the genetic algorithm for feature selection. In this algorithm, an individual is represented as an array with a length equal to the total number of features in the dataset. Each element of the array corresponds to whether the feature is selected or not, with a value of 1 indicating selection and 0 indicating exclusion. For instance, if the second element of the individual is 1, it signifies that the second feature has been selected for inclusion in that specific individual. This representation allows for the formation of a population using randomly generated individuals. Each individual in the population represents a unique combination of selected features, with the goal of optimizing the selection process to identify the most effective features for the task at hand.

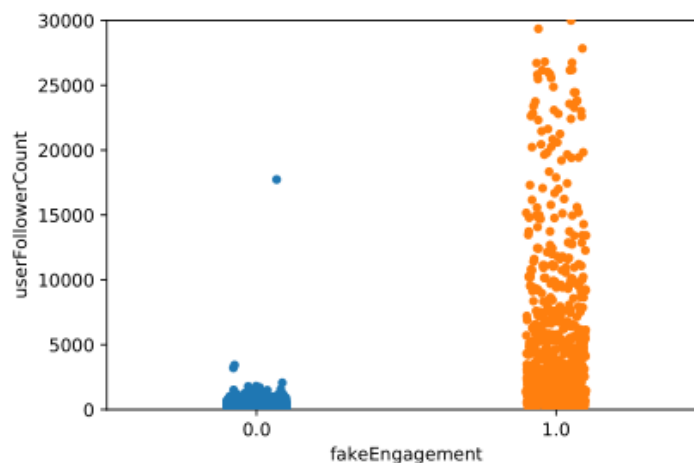


Fig 6: In-class data distributions for "follower count" feature.

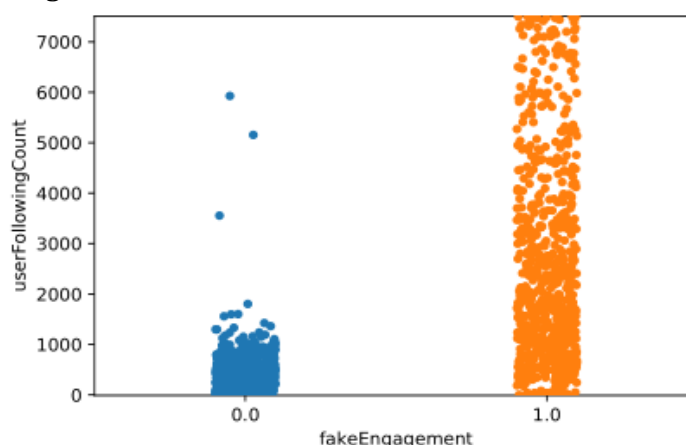


Fig 7: In-class data distributions.

$$\text{Fitness} = \text{F2 Score} - 2 \times \text{Tot.Feat.Cost} \quad (1)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

$$\text{F1 Score} = 2 \times \text{Recall} \times \text{Precision} / (\text{Recall} + \text{Precision}) \quad (4)$$

Fitness calculation formula is given in Eq. 1 Tot.Feat.Cost is calculated by summing the individual costs of the selected features. These costs are determined based on the reliability of the data collection which is discussed in the previous Bias Problem section. Realistic biases are represented with lower features costs while the negative bias is represented with higher costs. We conducted a comprehensive evaluation of several machine learning algorithms suitable for binary classification tasks. These algorithms included logistic regression, random forest, support vector machines (SVM), gradient boosting machines (GBM), random forest (RM), decision tree (DT) etc. Each algorithm was evaluated based on its performance in terms of accuracy, precision, recall, and F1 score using cross-validation techniques. Based on the results of the evaluation, we selected the algorithm that demonstrated the highest performance across multiple metrics while considering computational efficiency. Our decision was guided by the need to balance model accuracy with practical considerations such as deployment feasibility and resource requirements. The trained model was evaluated on the validation set to estimate its performance metrics, including accuracy, precision, recall & F1 score. The training process involved optimizing the model's parameters to achieve the best possible performance on the training data. We employed techniques such as grid search or random search for hyperparameter tuning, ensuring that the model generalized well to unseen data while avoiding overfitting.

IV. RESULT

By employing a variety of algorithms, the objective is to leverage different facets of the dataset, including independence, separability, and complex relationships, which have not been extensively explored in existing literature. The goal is to develop an effective method for detecting fake and automated Instagram accounts. To detect automated accounts, the selected features identified through cost-sensitive feature selection are utilized. In contrast, for identifying fake accounts, the base features from the fake-real dataset are directly employed.

In evaluating the effectiveness of the implemented techniques, Precision, Recall, and F1 Score are utilized as evaluation metrics, as outlined in equations 3-5, respectively. These metrics provide insights into the performance of the detection algorithms. Specifically, Precision measures the proportion of true automated accounts among all accounts identified as automated, while Recall calculates the proportion of true automated accounts correctly identified from all actual automated accounts. The F1 Score, which is the harmonic mean of Precision and Recall, offers a comprehensive assessment of the algorithms' performance by considering both false positives and false negatives. More meaningful metric for evaluating overall performance compared to Precision or Recall alone.

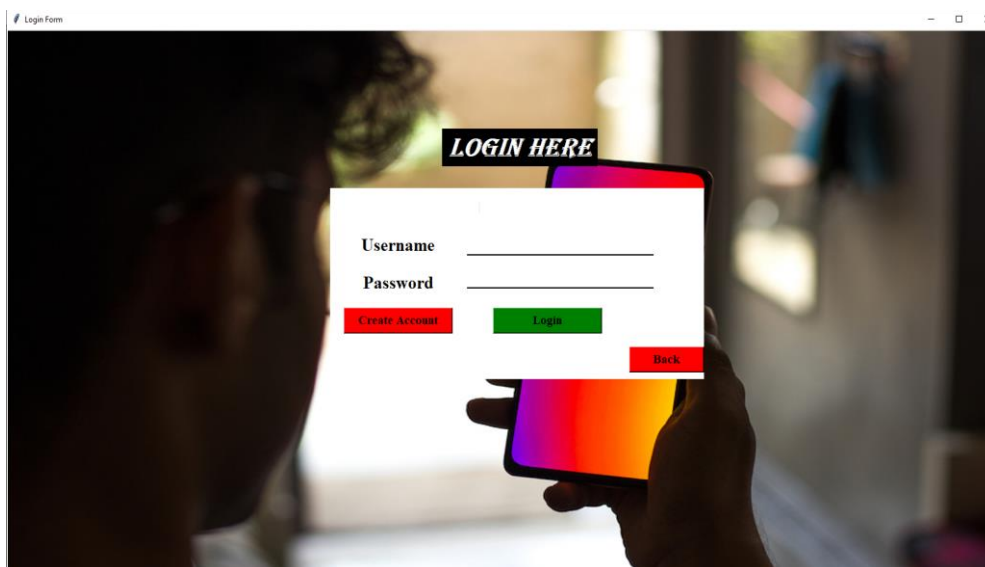


Fig 8: Login Page



Fig 9: GUI Main

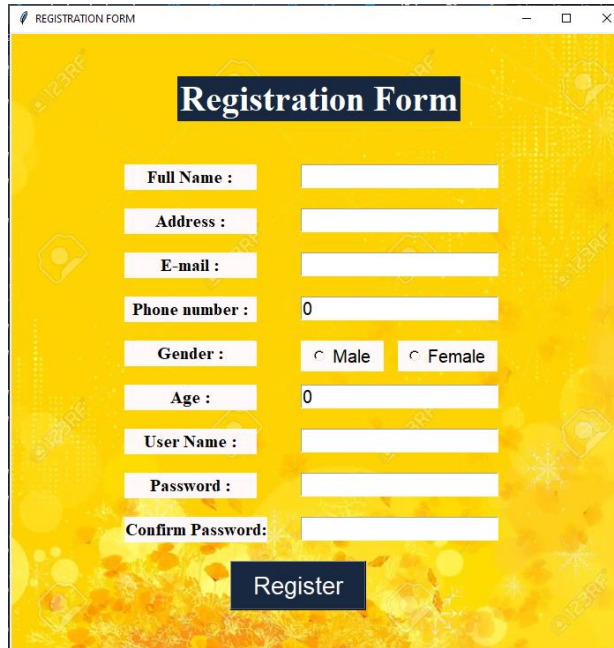
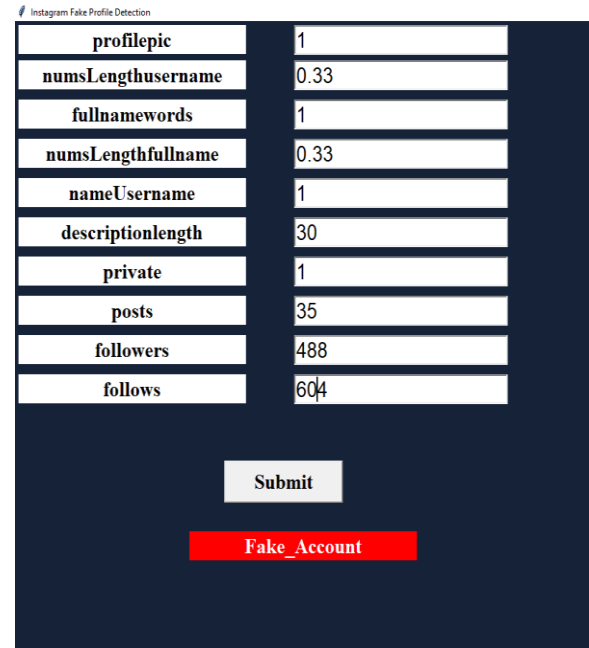


Fig 10: Registration Page



profilepic	1
numsLengthusername	0.33
fullnamewords	1
numsLengthfullname	0.33
nameUsername	1
descriptionlength	30
private	1
posts	35
followers	488
follows	604

Fig 11: Output Page 1



Fig 12: Output Page 2

V. CONCLUSION

The study titled "Fake Instagram Profiles Identification Using Machine Learning" offers a comprehensive strategy to address the persistent challenge of fraudulent profiles on social media platforms, particularly Instagram. Through the application of advanced machine learning techniques, this research endeavors to enhance the safety and credibility of the online environment, thereby fostering user trust and maintaining the integrity of social media communities.

The outcomes of this research transcend academic boundaries, exerting a tangible impact on individuals, businesses, and society at large. By leveraging the power of machine learning techniques, this research contributes to creating a safer and more trustworthy online environment for users, bolstering user confidence, and upholding the integrity of social media community. The research's outcomes extend beyond the realm of academia, impacting the lives of individuals, businesses, and society as a whole. As social media continues to shape the digital landscape, the work presented here contributes to building a foundation of trust and authenticity, reinforcing the positive potential of online interactions and collaborations.

VI. REFERENCES

- [1] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J Wisniewski, and Gianluca Stringhini 2022. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram. (2022), 1–14.
- [2] S. M. Din, R. Ramli and A. A. Bakar, "A Review on Trust Factors Affecting purchase Intention on Instagram", 2018 IEEE Conference on Application Information and Network Security (AINS), 2018.
- [3] Detect fake profiles on social media network.
- [4] S.C. Boerman, "The effects of the standardized Instagram disclosure for micro- and meso-influencers", Computers in Human Behavior, vol. 103, pp. 199-207, 2020.
- [5] S. Sheikhi, An Efficient Method for Detection of Fake Accounts on the Instagram Platform, 2020.
- [6] J. Kang and L. Wei, "Let me be at my funniest: Instagram users' motivations for using Fake insta", The Social Science Journal, 2019.
- [7] B. Mathew, R. Dutt, P. Goyal and A. Mukherjee, "Spread of Hate Speech in Online Social Media", Proceedings of the 10th ACM Conference on Web Science.
- [8] Y. Li, O. Martinez, X. Chen, J. E. Hopcroft, "In a world that counts: Clustering and detecting fake social engagement at scale,"
- [9] Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.
- [10] T. Clarke, "22+ Instagram Stats That Marketers Can't Ignore This Year".