# LIMFADD: LLM-enabled Instagram Multi-Class Fake Account Detection Dataset

Parthav Fulzele, Manu Mathew Jiss, Suksham Fulzele, and Tapadhir Das
Department of Computer Science, University of the Pacific, USA
Email: {p_fulzele, m_jiss, s_fulzele}@u.pacific.edu, tdas@pacific.edu

***Abstract*—Social networks like Instagram help connect billions of people. Unfortunately, this has resulted in cyber-criminals using social media platforms like Instagram to create fake accounts to target other users, like children and older adults. In this paper, we introduce the LLM-enabled Instagram Multi-Class Fake Account Detection Dataset (LIMFADD). Unlike state-of-the-art binary datasets that oversimplify the identification of fake accounts, LIMFADD enables classification into four distinct classes: real users, spam bots, scam accounts, and bot accounts. This allows us to gain additional insights into fake account behavior types on Instagram. Experimental results indicate an overall accuracy, macro-precision, macro-recall, and macro-F1 score of 0.97, 0.97, 0.97, and 0.97, respectively, across all classes. Additionally, we also achieve improved performance over state-of-the-art fake account detection datasets. Through this work, we aim to continue developing intelligent artificial intelligence solutions that can keep social media platforms like Instagram safer for users.**

***Index Terms*—Social Media, Fake Account Detection, Large Language Models, Instagram, Dataset**

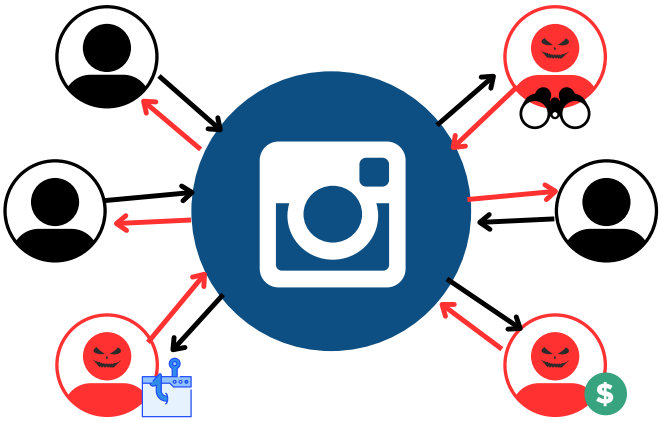

Fig. 1: Cybersecurity risks on Social Media Platforms like Instagram

## I. Introduction

Social networks have completely changed the way people talk and share media. Every day, billions of people use these platforms to connect. By 2025, around 5.42 billion people are expected to be on social networks [1]. Among these platforms, Instagram is a popular social media platform, with almost 2 billion users every month [2]. Due to this large user base, companies focus on social media marketing to reach customers, and the magnitude of this is expected to reach $276 billion in 2025 [3]. But despite popularity, many serious security problems have also been introduced in social media platforms like Instagram. Today, cyber-criminals use social media platforms like Instagram to create fake accounts to cheat other users, send spam messages, and even steal identities. The majority of these actions are targeted towards vulnerable portions of our population: children and older adults. According to the Federal Bureau of Investigation's Internet Crime Complaint Center, every year billions of dollars are lost due to online fraud [4]. In addition, the rise of social bots has made it harder to detect such activities [5], [6]. An illustration of this scenario is provided in Figure 1.

To combat these problems, artificial intelligence (AI) is being used to make platforms more secure. AI can analyze user behavior and quickly detect abnormal behavior patterns indicative of fake accounts. Platforms like Instagram are using these smart systems to ensure user safety and data protection. However, despite this development, the usage of AI in fake account detection for social media platforms currently still has limitations. Most state-of-the-art fake account detection datasets provide only binary classification capabilities, where AI models are trained to simply identify fake or real accounts. This encompasses all malicious behavior patterns into a single label, therefore limiting the insights that can be gained from multiple malicious behavior patterns. Two such datasets include the "Instagram Fake, Spammer, Genuine Accounts" dataset by Bakhshandeh et al. [7] and the "Instagram Fake and Real Accounts" dataset by Jafari et al. [8]. Additionally, due to the evolving tactics employed by fake accounts on social networks, it is difficult for static data sets to keep up with these behavioral changes. This becomes especially difficult if AI models only have a binary view of the behavior, as currently possessed by the existing datasets [9].

To address these limitations, in this paper, we introduce a new dataset called the Large Language Model-enabled Instagram Multi-Class Fake Account Detection Dataset (LIMFADD). This dataset is unique from the state-of-the-art datasets in existence as it does not simply label accounts as fake or real; it classifies fake accounts into multiple categories: spam bots, fake followers, and malicious promoters. This allows us to gain unique insights into individual types of malicious activity and their detection on Instagram. Using the generated dataset, we analyze its performance in detecting fake and real Instagram users in a multi-class situation. Also, the dataset harnesses the ability of large language models (LLMs)

to generate dataset features that are similar in magnitude to the intended feature spaces of normal accounts, spam bots, fake followers, and malicious promoters. This provides us with a wider dataset to train AI models. Additionally, we also evaluate the dataset using Explainable AI (XAI) to identify the dataset features that are influential in determining the final label for each account type. To the best of our knowledge, this is the first multi-class classification dataset for fake account detection for Instagram. Our main contributions include:

1) Developing the LIMFADD dataset for multi-classification in fake accounts on Instagram.
2) Evaluating dataset performance using standard performance metrics.
3) Evaluating dataset performance using XAI to increase transparency in predictions.

The rest of the paper is structured as follows: Section II introduces background work on fake account detection. Section III discusses our methodology in this paper, while Section IV presents the experimentation setup and results. Finally, conclusions are drawn and future work is provided in Section V.

## II. Related Works

Fake account detection in social media profiles such as Instagram is a trending research topic in user security and privacy. Besides traditional methods such as user feedback and manual inspection, there has been tremendous stress placed on automated methods capable of detecting fake accounts. Some works have investigated the usage of rule-based mechanisms to detect fake Instagram accounts [10] [11]. Authors in [10] developed a set of if-then rules to categorize Instagram profiles according to their legitimacy. Similarly, researchers in [11] developed the Knowledge Rules-based model to extract knowledge rules necessary for identifying fake accounts. The limitation of rule-based methods is that they are rigid in their pre-defined nature and can be circumvented by fake profiles that intentionally bypass those rules.

Researchers also introduced AI-based mechanisms for fake account detection, like user facial recognition [12] [13] [14]. In [12], the authors introduced face-processing methods as a double-factor authentication step in detecting fake accounts on social media platforms. Researchers in [13] designed and trained a neural network model and proposed a new algorithm to perform automated fake account detection using facial recognition. Work in [14] investigated the usage of convolutional neural networks towards facial recognition for fake account detection in social media profiles. The limitation of using facial recognition methods in this problem space is that they rely on the availability of profile photos to make decisions. On Instagram, when users are in private mode, only their profile photo is visible to other non-connected users. This photo can sometimes remain blank, which can impact the efficiency of facial recognition-based methods.

Researchers have also looked at AI with tabular data from Instagram profiles to discern between real and fake accounts [15] [16] [17] [18]. Authors in [15] introduced a Long Short-Term Memory (LSTM) network to classify fake and real accounts on Instagram. The work in [16] introduced Binary Grey Wolf Optimization and Particle Swarm Optimization algorithms to perform fake account detection on Instagram. Researchers in [17] introduced AI algorithms to perform fake account detection on Instagram. Finally, the work in [18] introduced a non-dominated sorting genetic algorithm for feature selection to detect fake accounts. While these works provide insightful results, the majority of them utilize binary classification datasets like [7] and [8]. Binary classification limits the insights that can be gained, as all malicious activity types are categorized under one label. In contrast, our proposed LIMFADD dataset provides the opportunity for multi-class classification which can provide additional insights to the detection process.

## III. Methodology

The methodology for developing our proposed LIMFADD dataset is provided in Figure 2.

### A. Real Account Data

To ensure a credible baseline, we gathered data from real Instagram accounts. These accounts were carefully chosen by our team and mainly contained the main profiles of known contacts such as friends, peers, co-workers, and family members. This data showed a true representation of how a user behaves by possessing natural follower-follower ratios, authentic posts, and a corresponding level of activity. These accounts were the baseline for comparing and segregating the other fake behaviors. Let the set of real account samples be denoted by $X_R$ with a total number of $A$ samples. Each sample was labeled *"Real"*.

### B. Spam Account Data

Spam accounts were identified through repetitive text observation in comments that were sent to many users by automated bots. These profiles regularly posted messages that appeared to be the same or highly similar to those found on different users' posts, promoting either their external links or their services. Such comments were compiled from the manual scanning of comment sections and direct message requests. These comments displayed the fatal hallmarks of spamming behavior, such as high-frequency messaging and abnormal engagement spikes. Let the set of spam account samples be denoted by $X_S$ with a total number of $B$ samples. Each sample was labeled *"Spam"*.

### C. Scam Account Data

Scammer accounts usually tend to trick users with phishing links or fraudulent promotions. Frequently, these accounts would have suspicious URLs in their bios or comment sections, with the coaching methods that these accounts would use to lure users. We manually scoured the comment threads and searched for profiles that followed such dubious tactics as fake giveaways or impersonation strategies. Let the set of

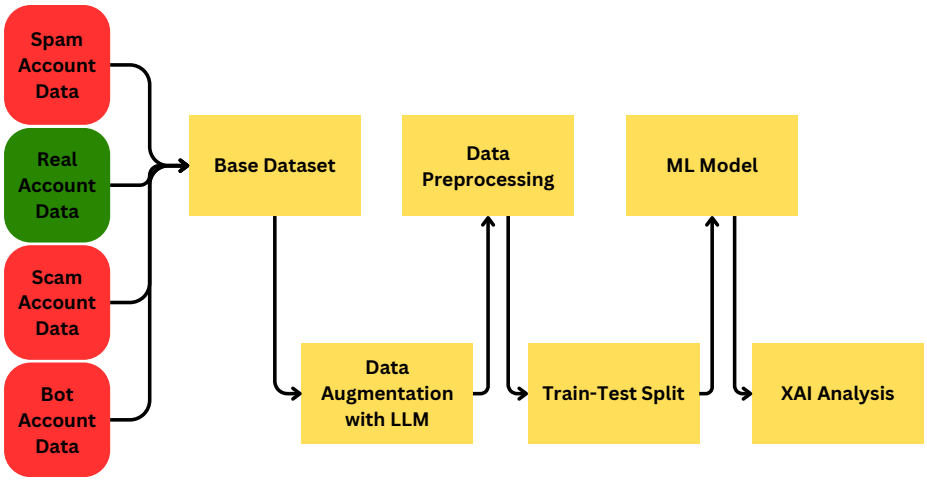


Fig. 2: Methodology for developing LIMFADD dataset

scam account samples be denoted by $X_K$ with a total number of $C$ samples. Each sample was labeled *"Scam"*.

### D. Data from Bot Accounts

Bot accounts are separated from regular accounts by the behavior they exhibit: liking, commenting, and following at very high frequency. Some bots behave unnaturally, such as commenting on generic phrases on many posts or having extremely high numbers of 'following' compared to followers. Most of these were found through our inbox or activity feeds, tagged manually. We classified the patterns that signaled automation in our dataset. Let the set of bot account samples be denoted by $X_B$ with a total number of $D$ samples. Each sample was labeled *"Bot"*.

### E. Base Dataset

After gathering initial examples from all four categories: real, spam, scam, and bot, we constructed a base dataset by equally selecting accounts from each class, i.e., $A = B = C = D$. This balanced selection ensured that no single class dominated the model's training process, thereby limiting the impact of any class imbalance on future AI models. The list of features we captured for all labels is provided in Table I.

### F. Data Augmentation using LLM

This initial base dataset was not substantial due to the manual annotation. However, it provided us with a stable baseline to further generate synthetic data samples whose magnitudes matched the feature spaces that were already established in our base dataset. To increase the size of the dataset using the established feature spaces, we utilized Chat-GPT [19] to assist with further data generation. Using this LLM, we increased the size of all four datasets $X_R, X_S, X_K$, and $X_B$ to $T$ samples each, where $A < T$, $B < T$, $C < T$, and $D < T$. By leveraging LLMs like Chat-GPT, we were able to expand our datasets from a relatively small size to a significantly more developed version. This allows for additional insights into a multi-class classification for fake Instagram accounts.

TABLE I: List of Features in LIMFADD

| **Feature** | **Definition** |
|---|---|
| Followers (number) | Number of followers on account |
| Following (number) | Number of users following |
| Follower/Following (number) | Ratio of Followers to Following |
| Posts (number) | Number of Posts |
| Posts/Followers (number) | Ratio of Posts to Followers |
| Bio (Y/N) | Does the user have a biography? |
| Profile Picture (Y/N) | Does the user have a profile picture? |
| External Link (Y/N) | Does the user have an external link on the profile? |
| Mutual Friends (number) | Number of mutual friends with the account |
| Threads (Y/N) | Does the user have Threads posts? |

### G. Data Preprocessing

After augmenting the final dataset, we performed data preprocessing. Here, our dataset was cleaned, and standardization was done. Features such as the "followers/following" ratio were duly examined, but sometimes ended with divide-by-zero errors. In such cases, we set those feature magnitudes to "0" to maintain integrity. Finally, concerning class labels, we encoded them numerically: "*Real* = 0", "*Spam* = 1", "*Scam* = 2", and "*Bot* = 3".

### H. Train-Test Split

The post-processed data was segmented into training and test sets based on an X/Y ratio split, where $X$ represented the number of training samples and $Y$ represented the number of testing samples. Feature scaling was done using normalization to ensure uniformity across numerical values. Model training was then performed on the training set while taking class representation on the set into consideration for the AI model to be able to discover meaningful patterns.

### I. Model

Since the main goal of the dataset is classification, any supervised model can be used for this task.

### J. XAI Analysis

To make the model predictions interpretable, we employed Local Interpretable Model-agnostic Explanations (LIME) as our framework, as it is a widely used explainable AI technique that indicates the features that contributed significantly to a specific prediction [20]. Using the LIME framework gave us additional insights regarding our dataset features that were most influential in label classification.

## IV. Results

### A. Setup

For the experiment setup, we used a Mac system powered by an Apple M1 chip with an 8-core CPU and 16 GB of RAM to run all experiments. The code was written in Python 3.10 using TensorFlow 2.x and Keras libraries for building the neural network. Jupyter Notebook was used for writing and testing the code, and we used Matplotlib and Seaborn for creating visualizations. For model training and evaluation, $X = 80$ and $Y = 20$, meaning that we used an 80-20 split between training and testing data. For experimentation, we set $A = B = C = D = 25$ and $T = 375$. Our total dataset had 1600 samples, with each label having 400 samples. For our model, we used a Deep Neural Network (DNN) due to its capability of learning various levels of representation from input data using multilayer structures [21]. Our model used ReLU activations in hidden layers and softmax in the output layer to handle multi-class classification. It was trained using Adam Optimizer with a learning rate $lr = 0.001$. The loss function used was categorical cross-entropy since it is a multi-class classification problem, and the model was trained for 25 epochs. For evaluation, we used Accuracy (A), macro-precision (P), macro-recall (R), and macro-F1 (F) scores.

### B. Results

First, we analyze the performance achieved by our DNN model on the proposed LIMFADD dataset under multi-class classification, as illustrated in Figure 3. We note that the model can generalize all labels across the dataset and achieves an A, P, R, and F score of 0.97, 0.97, 0.97, and 0.97, respectively. This highlights the impact of the LLM augmentation, which allowed us to have an equal representation of samples in the dataset to reduce class imbalance.

To provide further insights on the performance being achieved in the dataset under all the various classes, we also provide a heat map, as illustrated by Figure 4. We note that the model performs efficiently under all four labels of the dataset. It achieves an A, P, R, and F score of 0.97, 1.00, 0.98, and 0.99, respectively, when the label is "*Real* = 0." Similarly, it achieves an A, P, R, and F score of 0.97, 0.98, 1.00, and 0.99, respectively, when the label is "*Spam* = 1." Correspondingly, it achieves an A, P, R, and F score of 0.97, 0.95, 0.95, and 0.95, respectively, when the label is "*Scam* = 2." Finally, it

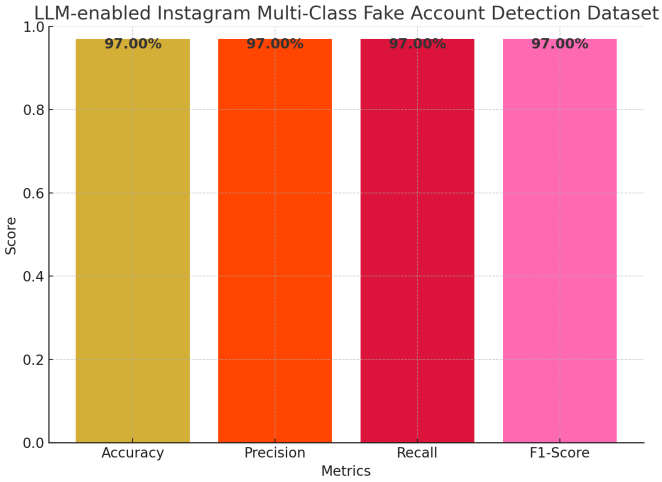


Fig. 3: Multi-class Performance Evaluation of LIMFADD Dataset

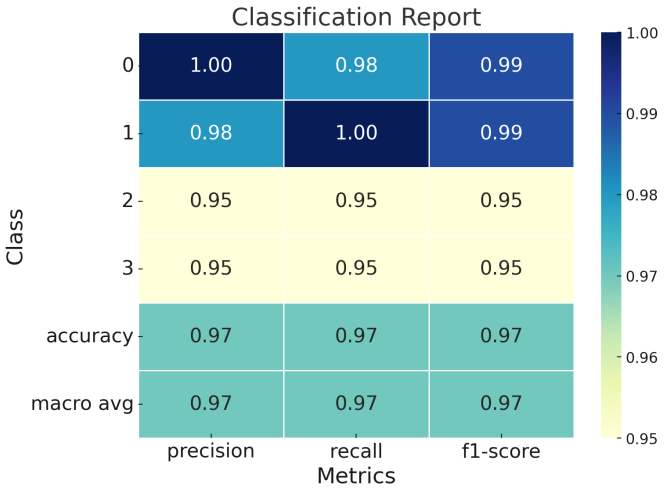


Fig. 4: Heat Map of LIMFADD Dataset Classes

achieves an A, P, R, and F score of 0.97, 0.95, 0.95, and 0.95, respectively, when the label is "*Bot* = 3." This yields an overall A, P, R, and F scores of 0.97, 0.97, 0.97, and 0.97, respectively.

Next, we also evaluate the performance achieved by the LIMFADD dataset when compared to state-of-the-art fake account detection datasets. In Figure 5, we provide a comparison between the LIMFADD dataset and the datasets in Bakhshandeh et al [7] and Jafari et al [8]. We note that under binary classification conditions, Bakhshandeh et al [7] achieve an A, P, R, and F score of 0.87, 0.86, 0.88, and 0.87, respectively. Similarly, Jafari et al [8] recorded even lower performance metrics with an A, P, R, and F score of 0.74, 0.89, 0.80, and 0.84, respectively. Our proposed LIMFADD dataset achieved the best binary classification scores out of the state-of-the-art fake account detection datasets with an A, P, R, and F score of 0.98, 0.96, 0.95, and 0.96, respectively.

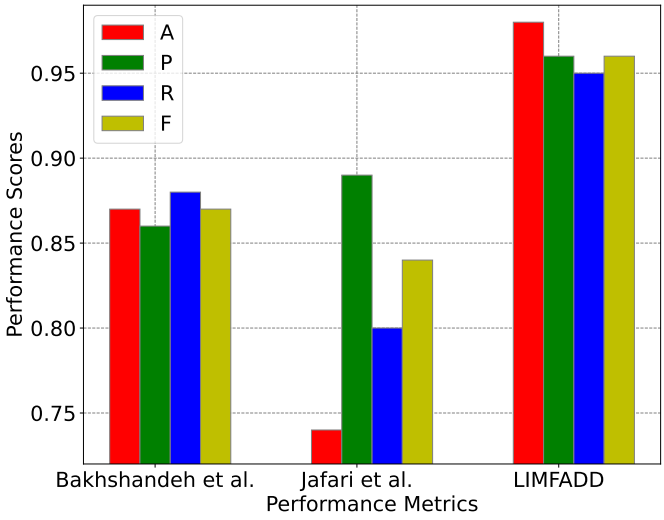


Fig. 5: Performance Evaluation between LIMFADD and other state-of-the-art datasets

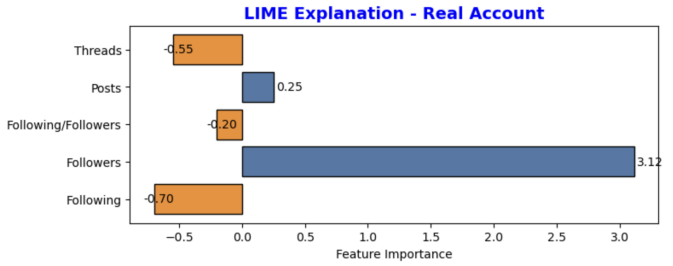


Fig. 6: LIME Explanation for Real Account

This highlights the importance of the LLM augmentation on the dataset, which allows us to effectively model both real and fake account instances. Additionally, a huge advantage of the LIMFADD dataset over the existing datasets is the ability to perform multi-class fake account detection. This allows us to gain new insights into the different types of fake accounts on Instagram and their different behavior patterns.

Finally, we illustrate the XAI analysis of our dataset to highlight the influential features we are capturing that help the ML model distinguish between the labels. This is illustrated in Figures 6 and 7. In Figure 6, we analyze a data sample that was classified by the model as "Real." We note that the number of "Followers" was a key positive signal for real accounts. Additionally, having a correlated number of "Posts" to the number of followers also correlated to real account behavior. Correspondingly, we note that "Threads" and "Following" contributed negatively, indicating fake account-like behavior. Too many Threads posts or too many following accounts compared to followers was representative of malicious activity on Instagram.

Similarly, in Figure 7, we analyze a data sample that was classified by the model as "Spam." We note that the number of "Threads" posts compared to followers resulted in the model

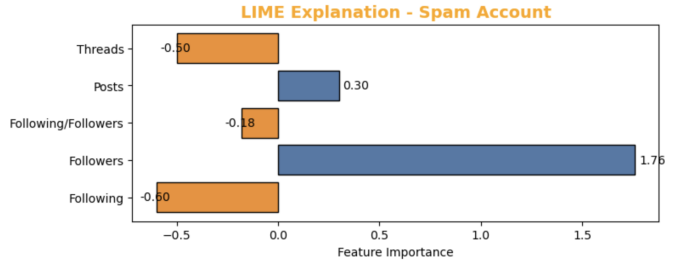


Fig. 7: LIME Explanation for Spam Account

labeling the sample as a fake account. Similarly, "Following" too many accounts without having an appropriate number of followers resulted in being classified as "Spam." On the flip side, "Followers" was a key positive signal for real accounts. This highlights that features like "Threads" and "Following" have strong negative contributions, representing patterns often associated with spammy or mass-generated accounts, which "Followers" is associated with real accounts on Instagram.

## V. Conclusion and Future Work

This work presents LIMFADD, a novel dataset that enhances the detection of social media fraud on Instagram through multi-class classification. Unlike state-of-the-art binary datasets that oversimplify the identification of fake accounts, LIMFADD enables classification into four distinct classes: real users, spam bots, scam accounts, and bot accounts. This allows us to gain additional insights into fake account behavior types on Instagram. The paper commences by introducing the impact of fraudulent activity on social media platforms like Instagram. We then describe the multi-class data pipeline for our dataset, including the data collection and LLM augmentation practices. Experimental results indicate an overall A, P, R, and F score of 0.97, 0.97, 0.97, and 0.97, respectively, across all classes. Additionally, we also achieve improved performance over state-of-the-art fake account detection datasets. In future work, the dataset will be expanded to include more content and behavioral features, such as post timing, interaction behavior, and natural language sentiment, to enable richer behavioral profiling. Additionally, we wish to expand this work to other social media platforms like Twitter/X, Facebook, and TikTok.

## References

[1] Statista, "Number of social network users worldwide," 2025. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] ——, "Number of monthly active instagram users." [Online]. Available: https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/

[3] ——, "Social media advertising spending worldwide." [Online]. Available: https://www.statista.com/statistics/237974/global-social-media-advertising-spending/

[4] FBI Internet Crime Complaint Center (IC3), "Internet crime report 2022." [Online]. Available: https://www.ic3.gov/Home/AnnualReports

[5] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016. [Online]. Available: https://cacm.acm.org/magazines/2016/7/204034-the-rise-of-social-bots/fulltext

[6] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, 2010, pp. 1–9.

[7] Kaggle user: free4ever1, "Instagram fake, spammer, genuine accounts," 2023. [Online]. Available: https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts

[8] Kaggle user: rezaunderfit, "Instagram fake and real accounts dataset," 2023. [Online]. Available: https://www.kaggle.com/datasets/rezaunderfit/instagram-fake-and-real-accounts-dataset

[9] K. Lee, B. 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. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2876

[10] G. Sonowal, V. Balaji, and N. Kumar, "A model to detect fake profile on instagram using rule-based approach," *CSI Transactions on ICT*, vol. 12, no. 4, pp. 95–105, 2024.

[11] S. Mohammed, N. Al-Aaraji, and A. Al-Saleh, "Knowledge rules-based decision tree classifier model for effective fake accounts detection in social networks." *International Journal of Safety & Security Engineering*, vol. 14, no. 4, 2024.

[12] A. Mughaid, I. Obeidat, S. AlZu'bi, E. A. Elsoud, A. Alnajjar, A. R. Alsoud, and L. Abualigah, "A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks," *Multimedia Tools and Applications*, vol. 82, no. 17, pp. 26 353–26 378, 2023.

[13] K. Kaushik, A. Bhardwaj, M. Kumar, S. K. Gupta, and A. Gupta, "A novel machine learning-based framework for detecting fake instagram profiles," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 28, p. e7349, 2022.

[14] V. Singh, R. Shanmugam, and S. Awasthi, "Preventing fake accounts on social media using face recognition based on convolutional neural network," in *Sustainable Communication Networks and Application: Proceedings of ICSCN 2020*. Springer, 2021, pp. 227–241.

[15] N. Alharbi, B. Alkalifah, G. Alqarawi, and M. A. Rassam, "Countering social media cybercrime using deep learning: Instagram fake accounts detection," *Future Internet*, vol. 16, no. 10, p. 367, 2024.

[16] P. Azami and K. Passi, "Detecting fake accounts on instagram using machine learning and hybrid optimization algorithms," *Algorithms*, vol. 17, no. 10, p. 425, 2024.

[17] S. Chelas, G. Routis, and I. Roussaki, "Detection of fake instagram accounts via machine learning techniques," *Computers*, vol. 13, no. 11, p. 296, 2024.

[18] A. Sallah, E. A. Abdellaoui Alaoui, A. Hessane, S. Agoujil, and A. Nayyar, "An efficient fake account identification in social media networks: Facebook and instagram using nsga-ii algorithm," *Neural Computing and Applications*, pp. 1–29, 2024.

[19] S. S. Biswas, "Role of chat gpt in public health," *Annals of biomedical engineering*, vol. 51, no. 5, pp. 868–869, 2023.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[21] S. Ahmed, T. ElGharbawi, M. Salah, and M. El-Mewafi, "Deep neural network for oil spill detection using sentinel-1 data: Application to egyptian coastal regions," *Geomatics, Natural Hazards and Risk*, vol. 14, no. 1, pp. 76–94, 2023.