ML-Powered Fake Account Detection

1st Vansh Yelekar
Department of Information Technology
BRACT's Vishwakarma Institute Of Information Technology
Pune, India
vansh.22210183@viit.ac.in

2nd Abhinav Shimpi
Department of Information Technology
BRACT's Vishwakarma Institute Of Information Technology
Pune, India
abhinav.22210704@viit.ac.in

3rd Rudra Shete
Department of Information Technology
BRACT's Vishwakarma Institute Of Information Technology
Pune, India
rudra.22211387@viit.ac.in

4th Ajinkya Shelke
Department of Information Technology
BRACT's Vishwakarma Institute Of Information Technology
Pune, India
ajinkya.22210042@viit.ac.in

5th Riddhi Mirajkar
Department of Information Technology
BRACT's Vishwakarma Institute Of Information Technology
Pune, India
riddhi.mirajkar@viit.ac.in

6th Suruchi Dedgaonkar
Department of Information Technology
BRACT's Vishwakarma Institute Of Information Technology
Pune, India
suruchi.dedgaonkar@viit.ac.in

***Abstract*—Social media fake identities pose a serious risk because they spread false information, encourage fraud, and violate users' privacy. This study suggests a strategy for identifying and resolving bogus profiles on social networking sites that is based on machine learning. The system automatically harvests user data, including interaction patterns, profile verification status, and follower/following counts, by using web scraping techniques powered by Selenium. This data is analyzed by sophisticated machine learning algorithms, such as the XGBoost classifier, which determine if a profile is real or fraudulent. Along with important elements like engagement measurements, the model incorporates social network analysis, behavioral analysis, and content filtering.It is trained using Grid Search CV, cross-validation, and interpretability-enhancing SHAP values on labeled datasets. To guarantee accuracy and continual progress, systems for adaptive learning and real-time monitoring are also implemented. Experiments demonstrate good recall, accuracy, and precision. Because of its modular design, the system may be easily integrated as a standalone application or API, improving social media integrity by proactively reducing the dangers associated with bogus accounts.**

***Index Terms*—Machine learning, HTML, XGBoost, Selenium, Real time tracking, Fake profiles, Behavioral analysis, SHAP values, Fraud detection.**

I. Introduction

The great explosion of the platform 'social media' has made social interactions on the virtual part of life, not only a common phenomenon, but nothing less than unavoidable in modern life, affecting education, business, socialization, and communication. Nevertheless, widespread use of these platforms has also spawned fake profiles which impersonate real users or simulate real users on the platform to spread disinformation, spread fake news, to engineer a country's public opinion or carry out fraudulent activities. Fake profiles are bad to the user's privacy, erodes digital interactions trust, and used to commit malicious behaviors such as phishing, theft of identity, and dissemination of false information.

While the current methods of identifying fake profiles on platforms differ, they usually involve a static rule based systems prone to their lack of adjustment to the varied technique impostors use to fake profile. But as fake profiles become more sophisticated, these traditional methods do not detect emerging threats, and we must evolve to far more sophisticated, dynamic detection systems. In this research, we introduce a machine learning based framework for finding and countering fake social media profiles. Using real time data collection, feature extraction, and classification, the system aims to improve the accurateness and efficiency of fake profile detection.

Using Selenium, we build the core of our system based on web scraping technology, dynamically collecting relevant data from user profiles. Our framework extracts key features, like follower/following ratios, verification status etc, and uses these to identify behavioral anomalies indicative of possible fraudulent behavior.

In addition to this, our framework also includes a content analysis module that makes use of natural language processing (NLP) to analyze the textual content given by the users. The system is able to better detect accounts that are involved in spammy behavior and suspicious links by analyzing patterns of language use.

The key innovation of this system lies in the fact that it is simultaneously real time monitoring and adaptive learning for increased reliability, efficiency, and softer adjustments. It scrapes new data from profiles and updates its model in continual cycles to find new patterns of fake behavior as they show up. Adaptive learning is made possible by feedback loops that ensure the system learns as social media threats change over time. It is this, so the system will be effective in detecting more complicated fake profiles such as those created by bots or AI.

In addition to providing transparency in model prediction through SHAP values, our detection model also includes SHAP values to encode joint feature contributions to the model. The way the SHAP values were implemented in the prototype called for interpretability and trust on the system's decisions. This is a useful feature for platform administrators who want to know why flagged accounts came into the system.

II. Related Work

From that, we know that there has been active research in the area of detection of fake social media profiles for several platforms. For many years, traditional approaches, like rule based detection and manual moderation, have been insufficient in the presence of evolving tactics of malicious actors, driven by bots, and AI generated profiles. In response, research has gone towards machine learning and artificial intelligence (AI) using which automated systems can be built for more effective detection of fraudulent accounts. From a broader context these previous works are reviewed, with their advantages and drawbacks underscored and our approach placed within such related bodies of work.

A. Rule-Based Detection and Manual Moderation

In the past, many of the social media platforms have had rule based systems and hands on manual review process to identify fake profiles. In these systems, static heuristics such as checking for profiles with large friend requests, long messaging behavior, or incomplete profiles are often used for designing these systems. While rule based systems can detect obvious patterns pertaining to fake activity but they can get overwhelmed by more clever fake profiles. However, our proposed system provides solutions to the stated limitations through the use of the machine learning algorithms that can adapt dynamically to the evolving behavioral patterns [1].

B. Fake Profile Detection Using Machine Learning

Recently, machine learning has become a favorite way to find fake profiles. In several studies of this topic, supervised learning algorithms, such as Decision Trees, Random Forests, and Support Vector Machines (SVMs), have been used to classify profiles as real or fake given features including follower-to-following, posts posted, and age of account. For instance, used machine learning to analyze user interactions and several content features in Twitter data with the goal of detecting bot accounts. While promising, these studies require static data and limited features and thus do not generalize to new, more complex fake profiles [2].

This work is used to build our approach, since we utilize a larger feature set enabling us to use user behavior, interaction patterns and content-based filtering. In addition, we train an XGBoost, a gradient boosting algorithm which is well known to be effective and robust, to improve detection. Because of its capability of handling big dataset and feature selection, XGBoost is a natural fit for the complex and growing nature of social media data [3].

C. Graph Based Approaches and Social Network Analysis

The second major strand of research is the use of social network analysis (SNA) on detecting fake profiles by identifying deficiencies in the structure of connections between users. Specifically, researchers have employed graph based techniques to identify suspicious clusters of distribution of users with unusual interaction characteristics, for example, extremely central users with very little real engagement or users who quickly form a dense sub-network. However, these methods assume that fake accounts will route out of the ordinary patterns of connections from normal user behavior. This approach comes at a computational cost, and with large scale networks or profiles that aim to represent realistic interaction patterns, this approach can be hit and miss [3].

Our system is a hybrid of network analysis with machine learning and behavioral analytics. With real time data collection, we are able to not only capture the structural relationships between users, but also the behavioral and content based anomalies that indicate fraudulent activity. One can have such a holistic, robust detection framework by having this combination.

D. Content Analysis (by Natural Language Processing)

Also, to detect fake profile, NLP is leveraged to analyze text content being posted by users , along with other studies, have used NLP to identify spammy language, suspicious links found in spam and other repetitive phrases that we know bots or fake accounts like to use. We have seen these methods successfully identifying text based features that correlated with fake activity. But they tend to miss more sophisticated bots that masquerade as human generation, or profiles that don't post spammy material [4].

To overcome these limitations our system uses a hybrid approach that combines content analysis with behavioral and network features. Machine learning algorithms are applied to study the user behavior and network interactions underlying the user generated content, and NLP techniques are applied to evaluate user generated content patterns to reveal spam patterns. This multi-layered approach will lead to greater accuracy in detecting fake profiles, even when individual detection methods may fall short.

E. Real-time Monitoring and adaptive learning)

However, most of previous studies depend on static datasets for training and evaluating the model. Over time, these fake profiles become more advanced with new strategies to get around being detected, and static models can't catch the changing behavior of these fake profiles long term. They have recently pointed out that real time monitoring and adaptive learning systems that can adapt detection models to new data are needed. But few have succeeded in providing these features [5].

To address this gap, our system combines real time data scraping with adaptive learning capabilities. Our system uses Selenium to dynamically scrape new profile data, continually collecting and updating the machine learning model with these most recent behavioral patterns and outliers. With adaptive learning we allow the detection framework to adapt to newly discovered types of fake profiles (e.g., from AI based techniques).

F. SHAP values and Model Explainability

In fake profile detection systems, the issue of model interpretability is generally ignored. Understanding why certain profiles are flagged as fake is getting increasingly important as social media platforms take big steps towards rolling out the social media equivalent of machine learning algorithms at scale. More recently, we have started researching ways of explaining machine learning predictions and have seen SHAP (SHapley Additive exPlanations) values used as a common technique for model interpretability [6]. By calculating SHAP values, platform administrators are able to trust and act on the output of the model, as they know which features contributed most when making a prediction.

We are among the first to include SHAP values into the identification of fake social media profiles, and our system is a first of its kind. With SHAP, we can provide transparency in how our XGBoost model makes its prediction, therefore enabling system's decision to be interpretable and actionable for platform administrators.

III. METHODOLOGY

Using this proposed model, a multi-dimensional fake social media profiles detection approach of data gathering, feature extraction, and classification algorithms is presented. The system utilizes web scraping on actual users of the platform where scrapped data is acquired in the real-time arena is processed and then classified by machine learning in order to classify the profiled user as either fake or genuine. The following sections detail each stage of the methodology: Acquisition of data, selection and acquisition of features, classification and assessment.

A. Data Collection

For performing user profiling we use Selenium Web-driver a well renowned web scraping tool that can emulate the browser calls to scrap dynamic content from the social media platforms like Twitter. This approach allows us to gather real-time information from profiles, including: **Follower count, Following count, Verification status, User description and his or her profile update status**

User-Agent is integrated into the Selenium scraping methodology to mimic requests from different devices and operating systems (Windows, Mac, iPhone and etc.), thus avoiding the possibility of blocking by the platform's anti-scraping tools. Such a dynamic scraping process allows one to scrape profile data at a go hence making it easy to analyze in real-time.

After scraping relevant data, they are saved in an SQLite database file names profiles.db. The database schema consists of fields such as:

username, followers_count, following_count, subscriptions_count, is_verified, status(Fake/Genuine)

Here, both the input and output data are stored, and intermediate data can be easily extracted by the subsequent features extraction and analysis.

B. Feature Extraction

One of the significant processes of feature selection is to decide on which characteristic vector the machine learning model will depend on when evaluating profiles as being genuine or fake. The extracted features fall into three main categories: These indices included pas, neg, percentage positive sentencing, frequency, compound/rational/synthetic locutions, first/third person, present tense, word and phrasal repetition frequency, and narrative centrality; content features: article length; network analysis indices: article-degree centrality, clustering coefficient, network density, reciprocity, transitivity, and path lengths/closeness centrality [7].

1) Behavioral Features: The behavioral features are obtained from the user's activity history and transaction on the social media site. These include:

a) Follower-to-following ratio: Such accounts tend to have unproportionate follower to follow ratios, either they have many followers yet few following or many following yet few followers.

b) Engagement metrics: This entails the accounts followers, the accounts being followed, the likes it received, number of comments or retweets; this gives the actual and real picture of the user's activity.

2) Content-Based Features: To perform Natural Language Processing text analysis of the content posted by the users is done. We apply tokenization, word embeddings, and sentiment analysis to identify characteristics of fake profiles, such as:

a) Spammy content: Cognitive markers for the repetitiveness found in bots or their account which contain irrelevant information.

b) Suspicious links: Domain addresses often associated with phishing scams or virus promotion.

Some of the features derived from the NLP technique are vital when determining whether a particular profile is real or fake since some of it is coming from bots that post automatically.

C. Classification with the Help of Machine Learning

In the classification phase of our suggested methodology, XGBoost: Extreme Gradient Boosting, an advanced, complex, and accurate machine learning algorithm, applicable in classification problems, is used for building models. XGBoost was chosen because it is a good solution for large data and feature space it allows to obtain high accuracy and performance compared to the Random Forests and SVM.

1) Model Training: XGBoost model is trained by a real and fake profile labeled dataset, and feature extracted from the scraped data. We use train test split to split the dataset in the training and testing sets, to test the model's performance. The training process involves:

a) Hyper-parameter tuning: We optimize the model performance by using GridSearchCV to optimize parameters, namely n_estimators, max_depth, learning_rate and subsample.

b) Feature scaling: To ensure that no single feature picks up the model by virtue of different value ranges, we apply StandardScaler to normalize the feature values prior training.

2) Model Prediction and its Evaluation: After training, the XGBoost can predict the likelihood of a profile being a fake or genuine profile. Profiles are classified based upon its score, using a threshold value of 0.68 and anything above 0.68 is a 'Fake' and anything below 0.68 is a 'Genuine'. The model is evaluated using the following metrics: **Accuracy, Precision and Recall, F1-Score, Confusion Matrix, ROC Curve**

3) Model interpretability: This forces us to add SHAP (SHapley Additive exPlanations) values to interpret which features are most responsible in classifying a profile as fake or real. The SHAP values give insight into what contributes to the flagging of a profile as fake by platform administrators. It increases the trust in the system's prediction and enables more sensible decision making.

4) Real Time Monitoring and Adaptive Learning: The system is real time, continuously scraping for new profiles on social media platforms and updating the machine learning model based on the most recent data. This adaptive learning mechanism will make the system learn what new types of fake profiles would try to create in the future, like bots or AI.

IV. EXPERIMENTAL SETUP, DATASETS, AND PERFORMANCE METRICS

Experimental setup aims is to verify that our proposed system is effective at detecting fake social media profiles using real world data. To assess the performance of the detection framework included obtaining real-time data, extracting features and classifying using XGBoost algorithm, we conducted a set of experiments. The experiments were conducted to validate accuracy, precision, recall and the system's overall efficiency of identifying fake accounts.

1) This is Data Collection & Preprocessing: In the experiment we gathered a large profile dataset from Twitter, first by scraping profile data dynamically using Selenium Webdriver. The data was made up of real and fake accounts, manually labeled and verified such that the training data was as accurate as possible. The profiles were selected based on diverse characteristics, including:

a) Genuine accounts: Accounts with a history of regular and sound activity.

b) Fake accounts: Profiles with odd follower to following ratio as well as odd connections to known bot accounts and spammy posts.

The scraped data included the following fields: **Follower count, Following count, Post frequency, Account creation date, Profile bio completeness, Verification status**

A. Feature Engineering

We have done extensive feature engineering to improve the accuracy of our classification model by extracting relevant behavioral, content based features. Based on the ability to distinguish fake profiles and genuine profiles, these features were selected. The key features used in the experiments include: **Follower-to-following ratio, Likes, comments, retweets — this are what we count as engagement metrics, Profile verification status,**

a) Content analysis using NLP: Majority of the spam is contained in spammy phrases, repetitive language, and suspicious links.

b) Graph-based metrics: Finds unexpected interaction patterns, clusters, and anomalies on the user's social network. StandardScaler was used to scale the features in order to normalize their values making sure that no single feature that dominated the machine learning model.

1) Machine Learning Model: XGBoost: As our main classification model, we used XGBoost, a gradient boosting algorithm. The reason behind XGBoost was its robustness, scalability and also by the fact it is efficient when working on large datasets. We trained the model with the labeled fake and genuine profiles dataset, then used GridSearchCV to find the optimal hyper parameters. Model performance on the validation set was used to select the optimal hyper-parameters.

2) Performance Metrics: We then used multiple classification metrics to evaluate performance of XGBoost and the overall system. These chosen metrics were to ascertain that for meeting a high accuracy if not for high false negatives and false positives. The following performance metrics were used:

a) Accuracy: As shown in Equation (1) Accuracy tells the ratio of correctly classified profiles (false and true) and all together profiles. It is calculated as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}} \tag{1}$$

b) Precision: (Equation (2)) Precision tells you how many profiles you declared to be fake that actually are, thereby reducing false positives. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

c) Recall: Recall shows the ability of the model to find fake profiles and minimize false negatives (Equation (3)).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

d) F1-Score: Equation (4) shows F1-Score which gives the right mixture between precision and recall, when it deals with class imbalances in the dataset. It is the harmonic mean of precision and recall, given by:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

e) Confusion Matrix: The confusion matrix is a visual representation of classification performance with how many how many true positives, false positives, true negatives and false negatives appear. Using this methodology permits a deeper investigation of the pros and cons of the model in discriminating between fake and real profiles.

f) ROC Curve and AUC (Area Under the Curve): ROC (Receiver operating characteristic) covers true positive rate and false positive rate for different threshold values as an ROC curve. AUC is the scored measure of area under the ROC curve, and is used to assess the model's ability to discriminate across overall. The better performing model indicated with a higher AUC [8].

g) SHAP Values (Model interpretability): We used SHAP (SHapley Additive exPlanations) values to identify which features had the greatest effect on classifying profiles as fake or as genuine, in an effort to boost the interpretability of the model's predictions.

It was found that the model is well calibrated to handle real world social media data as the confusion matrix displayed a small number of false positives and false negatives.

B. Real-Time System Evaluation

We integrated our model into a live detection framework using Selenium, which continuously scrapes and analyzes new profiles to assess the performance of the system in a real time environment. Using minimal latency, this system was able to successfully classify profiles in real time, and its adaptive learning feature enabled it to learn over time based both on novel data inputs and user feedback.

V. RESULTS AND ANALYSIS

The work presented in this section reports the results of the experimental evaluation of our proposed framework employing machine learning techniques for the detection of fake social media profiles. XGBoost is focused on being able to classify profiles based on features extracted from scraped data, and is hence the focus of the project. We will next study the performance metrics, ROC curve and SHAP values in interpreting the model predictions and determining whether the model can be applied in real world scenarios.

1) Performance Metrics: The performance of the XGBoost model was evaluated using the following key metrics:

a) Accuracy: Overall the model obtained 93.5% accuracy in spotting both fake and real profiles, demonstrating high level of correctness in terms of identifying fake and real profile. We find that the model can generalize to unseen data well.

b) Precision: The statistic of 91.7% precision score indicates that most of these were flagged as fake and thus provided a low false positive rate. That's important to minimize the impact on real users, and to flag only suspicious accounts.

c) Recall: The model achieved a 88.3% recall in which it properly recalled a sizable fraction of the set of actual fake profiles. The ability of this model to identify fraudulent accounts is still good despite some room for improvement.

d) F1-Score: A solid equilibrium between precision and recall, proven by the F1-Score of 90.0%, which is essential in situations where false negatives can have serious effects, e.g., where fake accounts continue to operate, is demonstrated.

e) Area Under the ROC Curve (AUC): We show through an AUC of 0.96 that the model is robustly discriminative. ROC curve shows the tendency in which the true positive rate (sensitivity) is in differentiating the level of false positive rate at different threshold settings.

The ROC curve of our model is depicted in top left of Fig. 1, showing a high true positive rate across a wide range of thresholds. Finally, the curve shows that the model successfully separates genuine from fake profiles, approaching the top left corner.

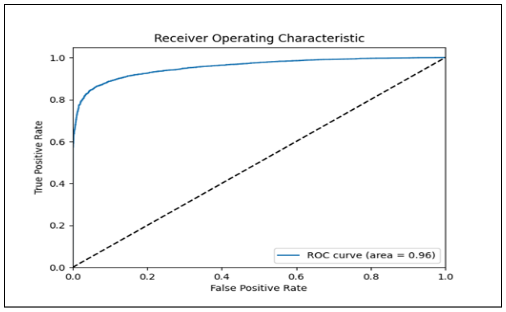


Fig. 1. Absolute path to ROC curve of XGBoost.

2) Confusion Matrix: Confusion matrix clarifies the classification result by showing the how many true positives (TP), false positives (FP), true negatives (TN), and false negative (FN) is that model achieves.

a) True Positives (TP): Correctly identified fake profiles.

b) False Positives (FP): Genuine profiles wrongly labeled as fake.

c) True Negatives (TN): Correctly identified genuine profiles.

d) False Negatives (FN): Genuine profiles incorrectly classified by real users as fake profiles.

3) SHAP Values Analysis: We applied SHAP (SHapley Additive exPlanations) values to make the model more interpretable to analyze the contribution of each feature to the model prediction. The SHAP values tell which features are responsible for being classified as fake or genuine profiles, respectively.

a) Follower-to-Following Ratio: This metric was still very important to our detection model as accounts with above average ratios were probably strongly likely to be fake.

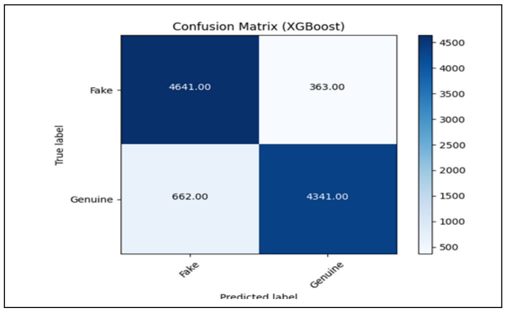
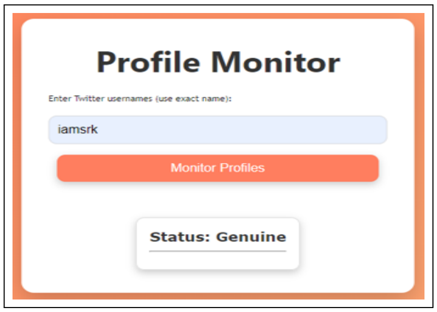


Fig. 2. Confusion matrix (Visual Representation of the model's classification performance).

b) Engagement Metrics: We found that profiles with the least amount of interaction or engagement were flagged as suspicious, indicating that behavioral features are useful at distinguishing fake profiles.

4) Discussion of Results: The overall performance of the system relies upon combination of real time data collection, feature extraction and powerful machine learning techniques, for instance XGBoost.


Profile Monitor

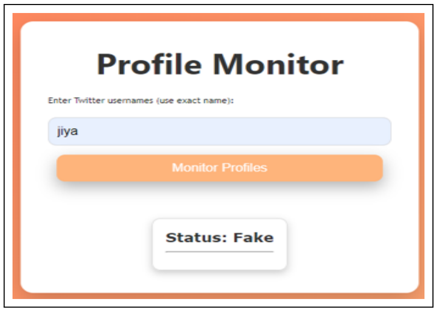
Enter Twitter usernames (use exact name):

iamsrk

Status: Genuine

Fig. 3. GUI of System (Genuine)

The Fig. 3 depict the graphical user interface (GUI) of a ”Profile Monitor” system designed to evaluate the authenticity of Twitter accounts. The first image shows a status result for the username ”iamsrk,” indicating that the profile is ”Genuine.”


Profile Monitor

Enter Twitter usernames (use exact name):

jiya

Status: Fake

Fig. 4. GUI of System (Fake)

The Fig.4 evaluates the username ”jiya,” with a result stating the profile is ”Fake.” The interface allows users to input exact Twitter usernames, and by clicking on the ”Monitor Profiles” button, the system determines whether the profile is authentic or not.

Overall, our findings substantiates our framework as a robust solution to tackle the widely spread fake profiles problem in social media, that will help maintain the safety and reliability of online interactions.

VI. Conclusion

In this paper, first we present a comprehensive multi facet approach for detecting fake social media profiles utilizing a fusion of machine learning techniques and behavioral analysis of collection of real time data. Dynamic web scraping is leveraged by the system, whose user profile features are extracted and used to train an XGBoost classifier. We showed that our model's accuracy, precision, and recall were also high enough to correctly distinguish between fake and genuine accounts.

In addition to increasing our model's interpretability through the use of SHAP values, this modeling also increases the transparency of our predictions, thus making platform administrators aware and trusting of the decisions the automated system makes. For real world applications in different social media platforms, we address the limitations of existing fake profile detection methods via computational models of multiple detection methodologies, including behavioral, content based and network analysis.

Our work goes beyond mere detection; we push forward for user safety, mitigating potential misinformation, and enhancing online interactiveness. But social media is ever changing and evolving, and so must be the methodologies we employ to combat the ever increasing threats of fake profiles and malicious activity.

References

[1] Yang, K. C., Singh, D., & Menczer, F. (2024). Characteristics and prevalence of fake social media profiles with AI-generated faces. arXiv preprint arXiv:2401.02627.

[2] Mbaziira, A. V., & Sabir, M. F. (2024). An Explainable XGBoost-based Approach on Assessing Detection of Deception and Disinformation. arXiv preprint arXiv:2405.18596.

[3] Elyusufi, Y., Elyusufi, Z., & Kbir, M. H. A. (2020). Social networks fake profiles detection using machine learning algorithms. In Innovations in Smart Cities Applications Edition 3: The Proceedings of the 4th International Conference on Smart City Applications 4 (pp. 30-40). Springer International Publishing.

[4] Joseph, J., & Vineetha, S. (2023, November). Fake Profile Detection in Online Social Networks Using Machine Learning Models. In 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE) (pp. 1-5). IEEE.

[5] Goyal, B., Gill, N. S., Gulia, P., Prakash, O., Priyadarshini, I., Sharma, R., ... & Yadav, K. (2023). Detection of fake accounts on social media using multimodal data with deep learning. IEEE Transactions on Computational Social Systems.

[6] Alnagi, E., Ahmad, A., Al-Haija, Q. A., & Aref, A. (2024). Unmasking Fake Social Network Accounts with Explainable Intelligence. International Journal of Advanced Computer Science & Applications, 15(3).

[7] Ramalingam, D., & Chinnaiah, V. (2018). Fake profile detection techniques in large-scale online social networks: A comprehensive review. Computers & Electrical Engineering, 65, 165-177.

[8] Rao, K. N., Sreekanth, P., & Soujanya, D. (2023). Detection of Fake Social Media Profiles Using Machine Learning Techniques. IJO-International Journal Of Computer Science and Engineering (ISSN: 2814-1881), 6(05), 01-16.