electronics

MDPI

Article

Advanced Algorithmic Approaches for Scam Profile Detection on Instagram

Biodoumoye George Bokolo * and Qingzhong Liu

Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, USA; qxl005@shsu.edu
* Correspondence: bgb023@shsu.edu

Abstract: Social media platforms like Instagram have become a haven for online scams, employing various deceptive tactics to exploit unsuspecting users. This paper investigates advanced algorithmic approaches to combat this growing threat. We explore various machine learning models for scam profile detection on Instagram. Our methodology involves collecting a comprehensive dataset from a trusted source and meticulously preprocessing the data for analysis. We then evaluate the effectiveness of a suite of machine learning algorithms, including decision trees, logistic regression, SVMs, and other ensemble methods. Each model's performance is measured using established metrics like accuracy, precision, recall, and F1-scores. Our findings indicate that ensemble methods, particularly random forest, XGBoost, and gradient boosting, outperform other models, achieving accuracy of 90%. The insights garnered from this study contribute significantly to the body of knowledge in social media forensics, offering practical implications for the development of automated tools to combat online deception.

Keywords: fake profile detection; machine learning; Instagram; ensemble learning; classification algorithms; feature engineering

Citation: Bokolo, B.G.; Liu, Q. Advanced Algorithmic Approaches for Scam Profile Detection on Instagram. *Electronics* **2024**, *13*, 1571. https://doi.org/10.3390/electronics13081571

Academic Editor: Aryya Gangopadhyay

Received: 28 February 2024
Revised: 15 April 2024
Accepted: 16 April 2024
Published: 19 April 2024

1. Introduction

Social media platforms have become seemingly ubiquitous in this day and age, fostering connections and enabling information sharing on a global scale. Platforms like Instagram have risen to prominence, shaping public discourse and becoming a crucial component of digital marketing strategies [1]. However, this interconnectedness also presents opportunities for malicious actors. One significant threat plaguing these platforms is the proliferation of online scams. Instagram, with its focus on visual content and user engagement, has become a particularly fertile ground for scammers, who employ deceptive tactics to exploit unsuspecting users [2].

These scams encompass a wide range of strategies, from phishing attempts to fake product promotions and the impersonation of legitimate accounts. The consequences can be severe, leading to financial losses, identity theft, and emotional distress for victims. Traditional methods of detecting scam profiles often rely on manual review by platform moderators, a time-consuming and inefficient approach that struggles to keep pace with the evolving tactics of scammers.

Traditional security measures often rely on user reports and manual verification, processes that are labor-intensive and not scalable [3]. This explains why, in many cases, even after identifying an Instagram account as fake or an impersonator, it often requires hundreds of user reports before any action is taken to resolve the issue. Machine learning offers a promising avenue to address this challenge. By leveraging sophisticated algorithms, we can automate the detection of scam profiles, enhancing user safety and creating a more trustworthy online environment. This study investigates the applicability of various machine learning models in identifying scam profiles on Instagram. We curate a comprehensive

dataset exceeding 65,000 profiles and meticulously preprocess the data to optimize them for model training and analysis.

Our study explores a wide variety of machine learning models, including decision trees, logistic regression, SVMs, random forest, K-nearest neighbors (KNN), XGBoost, gradient boosting, AdaBoost, extra trees, and a neural network implemented with Keras. We assess each model's performance using established metrics like accuracy, precision, recall, and F1-scores. Through this comprehensive analysis, we aim to pinpoint the most effective model for the detection of scam profiles on Instagram, paving the way for the emergence of robust and reliable detection systems that safeguard users from online scams.

This paper is meticulously structured to offer an in-depth exploration of the utilization of machine learning (ML) algorithms in detecting fake profiles on social media platforms, with a particular emphasis on Instagram. It begins with an introduction in Section 1, where the significance of the research, the challenges posed by fake profiles, and the potential of ML as a solution are presented. Following this, Section 2 delves into a comprehensive literature review, analyzing existing studies to provide a contextual backdrop and highlight the evolution of detection strategies in the field. The methodology is detailed in Section 3, explaining the data collection process, feature selection, and the rationale behind choosing specific ML algorithms, which lays the groundwork for the empirical analysis. In Section 4, the experimental setup and results are discussed, showcasing the metrics used for evaluation, as well as a comparative analysis of the performance of the selected suite of ML algorithms in identifying fake profiles. The discussion in Section 5 considers the findings' implications, acknowledges the study's limitations, and suggests future research directions, emphasizing the ongoing challenges in ensuring online security. The paper concludes in Section 6, summarizing the key insights, reflecting on the role of machine learning in enhancing social media security, and contemplating the impact of these advancements in fostering safer online environments.

2. Literature Review

Social networking sites are rapidly becoming a haven for online con artists, particularly Instagram, where it is easy to create false identities and target vulnerable individuals. The apparent ease of creating fake profiles, coupled with the popularity of Instagram, has led to an exponential rise in of scams and fraudulent activity on the platform. These scams have many forms, from financial schemes that prey on user confidence to phishing attacks that steal personal information. Resolving this issue requires effective detection techniques to identify and remove fake profiles. Traditional approaches often rely on manual review by platform moderators. However, this method suffers from significant limitations. The sheer volume of users and content on platforms like Instagram makes manual review a time-consuming and inefficient strategy. Instagram also implements this conventional approach—hiring human moderators to review profiles or any malicious content on the platform—an approach that is demonstrably ineffective as it sometimes requires hundreds of people to report a fake account before the platform flags the account and ultimately deletes it. As a result, many fraudulent profiles remain undetected, posing a continuous threat to user safety.

This literature review explores existing research that proposes methods and techniques to detect scams and fake profiles on social media platforms, mainly Instagram. This can help us to recognize the current advancements in the field and identify gaps that this research work can potentially fill.

Echoing Dwivedi et al.'s [4] description of social media's "dark side", where anonymity fuels online predation via financial scams and cyberbullying, fake profiles act as a digital wolf in sheep's clothing. Sahoo and Gupta [5] dissect the malicious intent behind these imposters, highlighting their use in elaborate financial schemes and information theft, often through phishing tactics. Furthermore, the research by [6] introduces the unsettling phenomenon of "social bots" or Sybils—fake profiles controlled by automated programs that mimic human behavior. Multiple studies [7,8] have documented these bots' nefarious

activities on social media, from spreading spam [9,10] and phishing links to manipulating unsuspecting users through deceptive tactics [11,12]. This pervasive threat underscores the urgent need for effective methods to identify and deactivate such fraudulent accounts.

In [13], Bharne and Bhaladhare suggest an approach that leverages a rich dataset of over 12,000 user profiles and their corresponding profile images to identify fake accounts within social networks. Their method involves a meticulous data preparation process, extracting features from both the textual content of profiles and the visual data within images. These features are fed into various machine learning models, including Naive Bayes, SVMs, and random forests (RF). Their research yields promising results, with the RF model achieving the best accuracy of 94.55%. This model also demonstrates a low false positive rate of 0.01, indicating a minimal likelihood of mistakenly flagging legitimate users. However, a false negative rate of 0.119 suggests room for improvement in catching all scammers. Interestingly, their most effective feature set incorporates n-grams, which capture sequential word patterns, and word2vec embeddings of size 10. This highlights the potential of combining textual and visual analysis for enhanced scam profile detection. Moreover, this work sets the tone for the performance superiority of ensemble methods over traditional classifiers. Hakimi et al. [14] focus on using traditional ML classifiers like KNN and SVM on a dataset of 800 facebook users. The models have 82.9% and 72.9% accuracy, respectively. These numbers are comparatively lower than those of other models surveyed in the field, especially considering that the model is trained and tested on a relatively small dataset. Harris et al.'s [15] work focuses on Instagram, using a suite of machine learning algorithms that contain both ensemble classifiers and traditional machine learning classifiers. Their work also suffers, not unlike many others in the field, from the very small size of their dataset—a mere 120 data points. However, their research results show that the two ensemble classifiers used in their work—random forest and XGBoost—return 100% accuracy, while KNN records 94.8%, SVM 85.34% and Naive Bayes 75%. These research works suggest the superiority of ensemble methods over traditional classifiers. However, since the generalizability of most of these works cannot be taken for granted due to the limited sizes of their training datasets, the need to leverage a substantial dataset to achieve robust and generalizable results then becomes apparent.

Other studies have explored the suitability of deep learning approaches—especially artificial neural networks (ANNs)—to tackle the task of fake profile detection on social media platforms. One study [16] addressed fake profile detection focusing specifically on Twitter. The researchers built a machine learning model to identify these fake accounts. Their dataset, harvested from the Twitter platform (now known as X), included roughly 2820 user profiles. To train their system, they explored five unsupervised learning algorithms, including SVMs and KNN. The researchers evaluated the system under two training and testing scenarios, a 70–30% split and an 80–20% split, incorporating four-fold cross-validation for robust assessment. Their findings revealed a trade-off between resource efficiency and accuracy. The 80–20% split offered lower resource consumption, while the 70–30% split yielded higher classification accuracy. Interestingly, when trained with the 70–30% split, the ANN algorithm achieved the most impressive performance with accuracy of 97.4%. Shreya et al. [17] worked on a Facebook dataset with 600 entries using RF, SVM, and ANN. Their results indicated that the ANN had the highest performance, with 96.73% accuracy, followed by RF with 92.33% and SVM with 88.67%. While research works that directly compare the performance of ensemble methods and neural network algorithms are limited, these two research works highlight the potential of deep learning approaches for this task, particularly for their ability to handle complex, non-linear relationships within the data. This then informs our decision to include neural networks (NNs) among the models in our suite of selected algorithms to tackle our research problem.

The correlation between spam and scams on social media platforms is significant, with spam often serving as a conduit for various scams. Spam includes unnecessary, irrelevant, or repetitive content that floods social media feeds or direct message inboxes. Scammers utilize tactics like creating fake accounts or hijacking existing ones to spread spam messages.

These messages range from fake remedies to romance scams, lottery fraud, and more. They exploit spam to deceive users, using sensational claims or fraudulent offers to entice individuals to click on links, share personal information, or fall for deceptive ploys. For instance, spam messages may contain links redirecting users to phishing websites or counterfeit products/services. Detecting and combating spam on social media platforms has become critical [18,19]. One such study that considered this task was that of Al-Zoubi et al. [20]. To investigate the characteristics of Twitter spam profiles, they assembled and examined a dataset of 82 user profiles. They applied and compared four classification algorithms in their work, namely decision trees, a multilayer perceptron (a popular artificial neural network model), KNN, and Naive Bayes. The Naive Bayes model had the best performance, with 95.7% accuracy. These results challenge the previously observed trend suggesting the inferiority of traditional classifiers like Naive Bayes and KNN compared to ensemble methods or neural networks.

Our research aims to build upon these foundational studies by not only integrating a broader spectrum of models, some of which have yet to be previously applied to Instagram's data, but also by providing a critical comparative analysis of these methods. The trends in the reviewed research have suggested the superiority of ensemble methods, the suitability of neural networks, and, in one outlying case, the ability of traditional classifiers to outperform other models. This informs our research design, aiming to select models comprising ensemble methods, traditional classifiers, and neural networks. We also choose to train and test these models with a significantly larger dataset compared to previous studies in this field. This is to ensure the viability and generalizability of our research findings. We delve into the specific algorithmic innovations and their limitations, highlight the adaptability of these algorithms to Instagram's unique environment, and discuss their resilience against the sophisticated evasion tactics of malevolent actors. By synthesizing these findings, our work seeks to fill the existing research gaps and suggest future directions for research, thus contributing to the ongoing refinement of the detection methodologies in the ever-evolving landscape of social media.

3. Methodology

3.1. Approach

To combat scam profiles on Instagram, we employ a comprehensive machine learning strategy—utilizing several machine learning algorithms. First, we collect a diverse dataset of user profiles from trusted sources like Kaggle, ensuring a balance between genuine and fraudulent accounts. After meticulous data cleaning and preparation, including handling inconsistencies and missing information, we explore feature engineering to enhance the profile details in the dataset. We then assess the effectiveness of our selected suite of machine learning algorithms in identifying scam profiles, considering each algorithm's strengths and weaknesses. We train them on our preprocessed data and evaluate their performance on a held-out test set using industry-standard performance evaluation metrics. This rigorous evaluation guides us in choosing the highest-performing model. Further analysis involves hyperparameter fine-tuning to optimize the performance. Ultimately, this systematic approach aims to identify the most effective machine learning model to mitigate scam profiles on Instagram, promoting a safer online environment for users.

3.2. Dataset Description

At the onset of this research work, there were two possible options available to us regarding data collection. We could collect data ourselves and build our dataset from scratch, or we could adopt well-curated datasets available online. Ultimately, we chose the latter, combining public data repositories for a dataset that was well suited to our work. The dataset was obtained from Kaggle—available at https://www.kaggle.com/datasets/krpurba/fakeauthentic-user-instagram (accessed on 4 March 2024). It was compiled by Purba et al. [21], who undertook similar work in identifying fake profiles on Instagram. They built the dataset by scraping Instagram data, capturing metadata and each user's

12 most recent media posts. The dataset comprises both genuine and counterfeit users, distinguished through human annotation. The genuine users were sourced from followers of select pages on Instagram, while counterfeit users were acquired by purchasing followers from Indonesian sellers.

The dataset contains 65,326 different user profiles—a very significant number, as most contemporary works on scam profile detection on social media platforms often use far smaller datasets. The dataset is balanced with a relatively equal number of records for each user profile class—32,866 fake and 32,460 genuine profiles. Seventeen distinct features were collected for each profile, encapsulating various aspects of a user profile, including the number of posts, profile picture presence, and engagement metrics. Each entry is labeled as 'fake' (f) or 'real' (r), indicating the class that it belongs to. Table 1 provides an in-depth overview of these features, detailing their descriptions and the rationale behind their inclusion in our analysis.

Table 1. Dataset feature overview and description.

Variable Name	Feature Name	Description	Rationale for Inclusion
POS	Number of Posts	The overall number of posts that the user has ever made	Fake accounts may have fewer posts
FLG	Number of Following	The number of Instagram accounts followed by the user	This helps us derive the following to follower ratio
FLR	Number of Followers	The number of accounts following the user	This helps us derive the following to follower ratio
BL	Biography Length	The length of the user's profile bio	Fake profiles often have shorter or no bios
PIC	Picture Availability	Binary value that indicates the presence or absence of a profile picture	Absence of a profile picture is more common in fake accounts
LIN	Link Availability	Binary value indicates an external URL's presence in the user profile	Fake profiles are less likely to link to external websites
CL	Average Caption Length	The typical character count for media captions	Presence of captions or lengthy captions are indicators of genuine accounts
CZ	Caption Zero	Percentage (0.0–1.0) of captions with nearly zero length (<=3)	The higher this percentage is, the lower the likelihood of genuineness
NI	Non-Image Percentage	Percentage (0.0–1.0) of media posts that are not a picture. An Instagram post can contain three different types of media: a picture, a video, and a carousel	Variety in the type of media posted by the user is an indicator of genuineness
ERL	Engagement Rate (Likes)	This is obtained by dividing the total number of likes by the number of media posts and the number of followers	Real accounts have higher engagement rates
ERC	Engagement Rate (Comments)	Same as ERL, but for comments	Real accounts have higher engagement rates
LT	Location Tag Percentage	Proportion of all media posts tagged with location (0.0–1.0)	Real users are more likely to tag their location on their posts
HC	Average Hashtag Count	The typical amount of hashtags used per post	This could indicate engagement activity or spam likelihood
PR	Promotional Keywords	The typical hashtag use of promotional keywords	Fake accounts may likely spam promotional keywords to boost their posts
FO	Followers Keywords	Average hashtag use for followers' hunting keywords, e.g., follow, like, follow back, f4f	Fake accounts may disproportionately spam follower keywords on their posts
CS	Cosine Similarity	The user's average cosine similarity between every pair of two posts	Useful in detecting spam accounts that repost identical or slightly modified content repeatedly
PI	Post Interval	Average interval between posts (in hours)	This could indicate the presence of automated activity or spamming behavior

The effectiveness of our model relies on the interplay between the various features extracted from user profiles on Instagram. These features, presented in Table 1, offer a multifaceted view into the user behavior and profile characteristics that can collectively distinguish between genuine and fraudulent accounts.

3.2.1. Content Analysis for Authencity

- Account Activity and Profile Completeness: Features like the number of posts (POS), follower count (FLR), and biography length (BL) paint a picture of the account activity

and effort invested in profile creation. A low number of posts, coupled with a large following (potentially inflated) and a short, generic biography, might indicate a hastily created, inauthentic profile used for scamming purposes.

- Content Consistency and Quality: The average caption length (CL) and the percentage of captions with minimal content (CZ) can reveal inconsistencies in communication styles. Scam profiles might resort to short, nonsensical captions or utilize automated tools generating irrelevant content. Additionally, the presence of a profile picture (PIC) and the distribution of media types (NI) can offer clues about legitimacy. Scam profiles might lack profile pictures or rely heavily on images for faster content creation, potentially sacrificing quality for quantity.

3.2.2. Engagement Patterns

- Engagement Metrics and Interaction: Engagement rates for likes (ERL) and comments (ERC) provide insights into user interaction levels. Scam profiles might have low engagement rates due to a lack of genuine followers or the use of automated bots to inflate their follower counts.

3.2.3. Hashtag Strategy and Deception

- Hashtag Usage and Content Relevance: The average number of hashtags used per post (HC) and the presence of specific keywords (PR, FO) associated with promotions or follower acquisition can indicate manipulative tactics. The excessive use of irrelevant hashtags or those aiming to attract followers through empty promises might be employed by scam profiles to gain visibility.

3.2.4. Identifying Unnatural Patterns and Content Theft

- Location Disclosure and Posting Consistency: The location tag percentage (LT) reveals a user's willingness to disclose their location. Scam profiles might avoid location tags to evade detection. The post interval (PI) reflects the user's posting habits. Sudden surges or inconsistent gaps between posts could suggest attempts to manipulate the feed or mask a lack of original content. Finally, the cosine similarity (CS) between a user's posts can expose potential plagiarism or the use of stolen content, often employed by scam profiles.

By analyzing this rich set of features and their interactions, our model can learn to identify patterns and red flags that deviate from the behavior of genuine users on Instagram. This broad spectrum of features enables a nuanced examination of profiles, allowing for the more accurate identification of fake accounts. Including diverse attributes ensures that our analysis does not rely solely on superficial indicators but delves into the depth of user engagement and presentation on the platform. The dataset's size, diversity in features, and clear labeling make it an apt choice for this study, providing a solid basis for the application and evaluation of various machine learning algorithms in detecting fake Instagram profiles.

3.3. Data Preprocessing

Data preprocessing is a critical phase in the data analysis pipeline, aimed at refining the dataset to ensure optimal utility for machine learning models. This phase encompasses meticulously executed steps designed to enhance the data quality and uniformity, laying a solid foundation for subsequent analytical endeavors. The preprocessing steps undertaken in our study include data cleaning, normalization, and encoding, each tailored to address specific aspects of data integrity and suitability.

- Cleaning: The foundation of our preprocessing effort was data cleaning, a process to identify inconsistencies within our dataset. We scrutinized each entry for missing values, outliers, or anomalies that could skew our analysis. This step does not merely concern exclusion; it seeks to preserve the integrity of the dataset, ensuring that the input to the models is of the highest quality and reliability.

- Normalization: Given the diverse nature of the numerical features within our dataset, normalization was an essential step to harmonize the scale of these variables. Numerical features, such as the total posts or number of accounts followed, inherently span a wide range of values, which, if left unadjusted, could disproportionately influence the model's learning process. To mitigate this, we applied normalization techniques to scale these features to a uniform range, typically between 0 and 1. This scaling ensures that each numerical feature contributes equally to the analytical models, preventing any single feature from dominating the learning process due to its scale. Normalization thus plays a pivotal role in fostering a balanced and equitable learning environment, where the influence of each feature is duly calibrated to its intrinsic significance rather than its scale.
- Encoding: The transformation of categorical features into a machine-readable format was accomplished through label encoding, an efficient technique that assigns a unique integer value to each category. This step is crucial in accommodating categorical data within machine learning algorithms that inherently require numerical input. This process was primarily performed on our target class attribute initially categorized as 'real' or 'fake'. Label encoding assigned the 'real' class the binary value of 0 and the 'fake' class the binary value of 1.

Together, these preprocessing steps constituted a comprehensive effort to refine the dataset, addressing potential sources of bias and variance that could compromise the efficacy of our machine learning models. By ensuring data cleanliness, uniformity, and compatibility with the algorithmic requirements, we set the stage for a rigorous process to unravel the complex dynamics of fake profile detection on Instagram.

3.4. Feature Selection and Engineering

The process of distinguishing authentic profiles from fraudulent ones on Instagram necessitates a discerning examination of various profile attributes to ascertain their predictive value. Our approach to feature selection and engineering was twofold, intertwining domain expertise with statistical methodologies to determine what constitutes genuineness in the vast volume of Instagram profiles. A key goal of our feature selection process was to obtain the most salient predictors from a broader set of attributes. This involved a meticulous analysis of each feature's distribution across genuine and fake profiles, seeking patterns or discrepancies that could signal underlying authenticity or a lack thereof. Domain knowledge played a pivotal role in this phase, guiding our scrutiny toward features that intuitively aligned with the behavioral norms of authentic users, versus the anomalies often associated with fraudulent accounts. The inclusion of a profile picture as a feature is based on the premise that genuine accounts are more likely to have a profile picture, whereas fake accounts might not, reflecting a lack of effort in ensuring profile completeness. The analysis extended to the lengths of usernames and full names and their ratios, under the hypothesis that fake accounts may exhibit atypical patterns in these aspects, diverging from the norms observed in genuine profiles. The bio description length is another considered feature, as genuine users tend to provide meaningful information in their bios, unlike fake profiles that might leave this section sparse or filled with irrelevant content. The privacy status of an account and the presence of an external URL were scrutinized, as these elements can offer insights into an account's legitimacy, albeit not being definitive indicators on their own. Engagement metrics, such as the total posts or number of accounts followed, are crucial in this analysis. They provide a quantitative measure of an account's social activity, where deviations from expected patterns could signal inauthentic behavior. The rationale for these feature selections is supported by the existing literature, which indicates that specific profile characteristics and behaviors are more prevalent among fake or inauthentic accounts. Through this meticulous feature selection process, our study aims to harness these insights, ensuring that each feature contributes significantly to the robust detection of scam profiles, thereby enhancing the accuracy and reliability of our machine learning models in identifying fraudulent activity on Instagram.

3.5. Machine Learning Models

In our study of the identification of fake profiles on Instagram, we meticulously selected a suite of machine learning algorithms, each chosen for its distinct advantages and appropriateness for the task. While we explored a comprehensive selection, some algorithms were excluded due to their better fit for specific data types. For example, our static profile features would not be best served by LSTM networks, which perform well with sequential data, such as time series. Similarly, algorithms like CNNs excel at image recognition, while RNNs and deep belief networks (DBNs) are often utilized for sequential or complex data structures. While valuable for dimensionality reduction, principal component analysis (PCA) and autoencoders are not prioritized in this case. Our focus lies in maintaining the interpretability of the features used for model decision-making. Therefore, we opted for algorithms that provide a clear understanding of how they classify data points.

With these considerations in mind, the suite of machine learning models chosen for this study included decision trees, logistic regression, SVMs, random forest, KNN, XGBoost, gradient boosting, AdaBoost, and extra trees. Our algorithmic strategy was designed to span straightforward to complex models, aiming for a thorough examination of their capabilities in detecting fake profiles while considering each model's specific strengths and limitations within the scope of bolstering social media security.

3.5.1. Decision Trees

Decision trees are widely used supervised learning tools that offer a tree-like structure where data points are classified based on sequential questions about their features. They are intuitive to interpret, as each split corresponds to a simple decision rule. This interpretability makes them valuable in understanding the decision-making process, making them suitable for our scam profile detection work. However, they can be susceptible to overfitting if not carefully pruned.

3.5.2. Logistic Regression

Logistic regression is a statistical technique used in binary classification tasks, where the output variable is categorical and can only have two possible outcomes (e.g., yes/no, or, in our case, legitimate or scam profiles). It assumes independence among variables, which is not always the case. However, its effectiveness and efficiency often overcome this statistical assumption. Its simplicity and interpretability also make it suitable for our work.

3.5.3. Support Vector Machines (SVMs)

SVMs operate by pinpointing the optimal hyperplane that effectively segregates data points into their respective classes. This hyperplane is crafted to maximize the margin in binary classification, representing the distance between the hyperplane and the closest support vector data points from each class. This model offers numerous advantages—such as consistent effectiveness in complex data spaces, the capability to manage non-linear decision boundaries, and resistance against outliers and overfitting, especially in scenarios where the dimensionality of features outweighs the sample size.

3.5.4. Random Forests

Random forests work by training multiple decision trees and generating predictions by averaging the forecasts of each tree for regression tasks or by selecting the mode of the classes for classification tasks. Compared to individual decision trees, the collective predictions from multiple trees in a random forest exhibit higher accuracy and reduced overfitting. Moreover, random forests demonstrate versatility in handling various types of features and exhibit resilience against noise. However, interpreting the inner workings of a random forest can be more challenging.

3.5.5. K-Nearest Neighbors (KNN)

KNN operates based on the principle of similarity, utilizing the average value or majority class of the k-nearest neighbors in the feature space (where k is an integer) to predict the label of a new data point. One of its advantages is its minimal training data requirement and ease of deployment. However, KNN may be susceptible to irrelevant features and heavily rely on the selection of the k parameter, which can significantly affect its performance.

3.5.6. XGBoost, Gradient Boosting, and AdaBoost

These are all ensemble boosting techniques that sequentially build models, where each new model learns to improve upon the errors of the previous one. They are powerful and flexible, often achieving high accuracy. However, they can be computationally expensive, and interpreting their inner workings can be complex.

3.5.7. Extra Trees

Extra (extremely randomized) trees are similar to random forests; this is an ensemble method that uses decision trees for classification. However, extra trees randomly split the features at each node instead of using the best split (like random forests or traditional decision trees). This reduces the risk of overfitting but can slightly decrease the accuracy compared to random forests.

3.6. Evaluation Metrics

To comprehensively assess the performance of our models, we employed a suite of evaluation metrics, each elucidated through mathematical formulations, to provide a multifaceted view of the model performance.

- Accuracy: Defined as the proportion of properly predicted occurrences (true negatives as well as true positives) to all instances in the dataset. It offers a straightforward measure of the overall model performance but may not be as informative in imbalanced datasets. The formula for accuracy is given by

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: Also referred to as the positive-predictive value, precision gauges how accurate positive predictions are. In situations where the cost of false positives is significant, it is especially crucial. The following is the precision formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

  Precision thus reflects the model's capability to identify only the relevant instances as positive.
- Recall (Sensitivity): This measure expresses how well the model detects all real positives, which is important when there is a high cost associated with missing a positive occurrence (false negative).

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score: By considering the harmonic mean of precision and recall, the F1-score unifies both metrics into a single measure. This measure is useful when finding a compromise between recall and precision is crucial, particularly when the class distributions are not uniform.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Experimental Setup and Results

A robust experimental setup is critical to validate machine learning models' performance. This section outlines the configuration of the experiment, including data handling, the computational resources, and the model parameters. For the training and evaluation of the models, we split the dataset into two subsets.

- Training Set: 70% of the data (45,728 profiles) were used to train the models. This subset provides the models with various examples of both fake and real profiles, enabling them to learn the underlying patterns associated with each class.
- Testing Set: The remaining 30% of the data (19,598 profiles) were held out for model validation. This separation ensures that the performance metrics reflect the models' ability to generalize to unseen data. The split was performed randomly to avoid any bias, but with stratification to ensure that the class distribution was proportionally consistent with the original dataset.

To ensure the reproducibility of our results and to provide a clear reference for the configurations of the machine learning models used, Table 2 presents a detailed breakdown of the hyperparameters and settings for each model. Table 2 encompasses the specific parameters used for models like K-nearest neighbors, extra trees, random forest, and XGBoost. The settings range from the choice of kernels in SVMs to the depth of trees in ensemble models. This detailed listing is pivotal in understanding the experimental setup and facilitates the replication of the study by other researchers in the field.

Table 2. Models, hyperparameters, and settings.

Model	Hyperparameters and Settings
XGBoost	use_label_encoder = False, eval_metric = 'logloss', StandardScaler in pipeline, StandardScaler in pipeline
Random Forest	n_estimators = 100, StandardScaler in pipeline
K-Nearest Neighbors	Default parameters (n_neighbors = 5), StandardScaler in pipeline
Neural Network (Keras)	Layers: Dense(64, activation = 'relu'), Dense(32, activation = 'relu'), Dense(1, activation = 'sigmoid'); optimizer = 'adam', loss = 'binary_crossentropy'; epochs = 30, batch_size = 32
Extra Trees	n_estimators = 100, StandardScaler in pipeline

The evaluation of each machine learning model in our study was based on key performance metrics, namely the accuracy, precision, recall, and F1-scores, on the test set. Table 3 presents a comprehensive overview of these metrics for each model utilized. Figure 1 also illustrates, graphically, the performance of each model across the chosen evaluation metrics. By showcasing the metrics side by side, both the table and the figure facilitates a comparative analysis, allowing for an assessment of the model's effectiveness in identifying scam profiles on Instagram. These metrics are pivotal in gauging the strengths and weaknesses of each model, providing valuable insights into their predictive capabilities and reliability. Such a comparative examination aids in the identification of the most appropriate models for scam profile detection, ultimately contributing to the development of robust detection systems.

Table 3. Performance metrics of machine learning models for scam profile detection.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	81%	82%	79%	80%
Decision Tree	86%	85%	86%	86%
Support Vector Machine	85%	88%	80%	84%

Table 3. *Cont.*

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	90%	95%	84%	89%
K-Nearest Neighbors	81%	83%	77%	80%
XGBoost	90%	94%	86%	90%
Gradient Boosting	90%	95%	84%	89%
AdaBoost	87%	87%	86%	86%
Extra Trees	88%	92%	83%	87%
Neural Network (Keras)	89%	93%	84%	88%

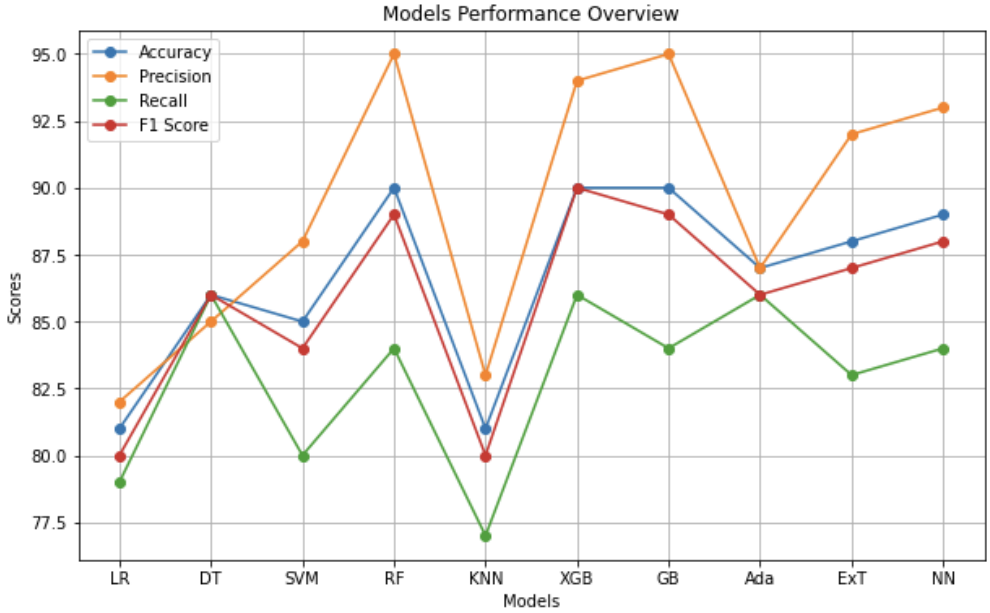

Figure 1. Model performance overview.

Key Findings

1. Ensemble methods, particularly random forest (90% accuracy, 95% precision, 84% recall, 89% F1-score) and XGBoost (90% accuracy, 94% precision, 86% recall, 90% F1-score), emerged as the frontrunners, exhibiting exceptional performance across all metrics. Their ability to leverage the strengths of multiple decision trees likely contributed to their superior ability to capture complex patterns within the data, leading to more accurate scam profile detection.
2. Gradient boosting (90% accuracy, 95% precision, 84% recall, 89% F1-score) and neural networks (Keras) (89% accuracy, 93% precision, 84% recall, 88% F1-score) also demonstrated promising results, achieving high accuracy and precision. These approaches offer flexibility in handling non-linear relationships within the data and might be further optimized through hyperparameter tuning.
3. Traditional models like logistic regression (81% accuracy), decision trees (86% accuracy), support vector machines (85% accuracy), KNN (81% accuracy), and AdaBoost (87% accuracy) provided a baseline performance level. While offering a reasonable degree of accuracy, they were surpassed by the ensemble methods and some advanced techniques.

The comparative analysis of the models indicated that the ensemble methods, particularly random forests, XGBoost, and gradient boosting, performed best across all metrics. Other ensemble methods like extra trees and AdaBoost were not far behind, with accuracy of 88% and 87%, respectively. These models benefit from aggregating the decisions of

multiple learners, which helps in reducing the variance and bias, leading to the better generalization of the test data. The logistic regression model served as an effective baseline, while the decision tree provided insights into the feature relationships. The support vector machine and K-nearest neighbors performed moderately well, suggesting the presence of non-linear decision boundaries.

While random forest and XGBoost achieved the highest overall performance, it is crucial to consider the trade-off between precision and recall when selecting the optimal model for real-world deployment. If minimizing the number of incorrectly flagged legitimate profiles is paramount, random forest (95% precision) and XGBoost (94% precision) might be preferable choices. For scenarios wherein a balance between catching a high proportion of scam profiles and avoiding false positives is desired, gradient boosting and neural networks represent strong options due to their balanced performance across metrics.

These results demonstrate the viability of ML models in detecting scam profiles on Instagram, with ensemble methods showing particular promise. These insights could be instrumental in the ongoing development of cybersecurity measures on social media platforms.

5. Discussion

The findings of this investigation illuminate the potential of various machine learning models in combating scam profiles on Instagram. The strong performance of the ensemble methods, particularly random forest (90% accuracy), aligns with observations made in related research. Bharne and Bhaladhare [13] achieved 94.55% accuracy when using a random forest model for fake account detection on social networks. Their work also highlights the value of combining textual and visual analysis, suggesting a potential avenue for future exploration in our own research.

Our results extend these findings by demonstrating the effectiveness of ensemble methods specifically for scam profile detection on Instagram. While Bharne and Bhaladhare [13] incorporated visual features like word2vec embeddings, our study focused solely on textual data extracted from Instagram profiles. This suggests that ensemble methods can achieve high accuracy even without visual analysis, potentially due to the rich information conveyed through textual content on Instagram profiles.

Another relevant study by Al-Zoubi et al. [20] explored spam profile detection on Twitter using various classification algorithms. Their findings, achieved with a dataset of 82 profiles, showcase the potential of Naive Bayes for this task (95.7% accuracy). However, our research, utilizing a significantly larger dataset of over 65,000 Instagram profiles, demonstrates the superiority of ensemble methods like random forest, XGBoost and gradient boosting for scam profile detection on a larger scale. This highlights the importance of the dataset size and the potential limitations of drawing conclusions from smaller datasets, as observed in Al-Zoubi et al.'s work [20].

The moderate performance of support vector machines (85% accuracy) and K-nearest neighbors (81% accuracy) in our study suggests that these models might not be as effective as ensemble methods for scam profile detection on Instagram. This aligns partially with the findings of Al-Zoubi et al. [20], who observed that Naive Bayes outperformed K-nearest neighbors for spam detection on Twitter.

The implications of this research are significant for cybersecurity and social media governance. The ability to automatically detect fake profiles with high accuracy can help in preemptively mitigating the risks associated with fraudulent activity on social platforms. Companies can integrate these machine learning models into their cybersecurity infrastructure to monitor and maintain the authenticity of user interactions.

Overall, this research contributes to the development of robust scam profile detection systems on Instagram by demonstrating the effectiveness of machine learning models, particularly ensemble methods. The findings provide valuable insights into the strengths and limitations of various models in this context, while also emphasizing the importance of the dataset size for generalizable results. By building upon these results and addressing

the limitations identified, future research could serve to develop even more accurate and generalizable detection systems, fostering a safer online environment for Instagram users.

Limitations

While this study sheds light on the effectiveness of various machine learning models in detecting scam profiles on Instagram, it is essential to acknowledge certain limitations. Firstly, the generalizability of our findings might be influenced by the specific characteristics of the chosen dataset. Future research could benefit from employing even larger and more diverse datasets to ensure the models' effectiveness across a broader spectrum of scam profiles. Secondly, this study focused on a particular set of machine learning models. Exploring a wider range of models, including more advanced deep learning architectures, could potentially lead to further improvements in the detection accuracy. Additionally, the research prioritized feature engineering based on readily available profile information. Investigating the creation of domain-specific features or incorporating visual content analysis through techniques like convolutional neural networks (CNNs) could offer avenues for further exploration. By acknowledging these limitations and pursuing future research directions, we can contribute to the ongoing development of robust and adaptable scam profile detection methods for social media platforms.

6. Conclusions and Future Work

Social media platforms like Instagram have become breeding grounds for online scams, posing a significant threat to user safety. This research investigated the effectiveness of various machine learning models in detecting scam profiles on Instagram. We employed a meticulously preprocessed dataset exceeding 65,000 profiles to train and evaluate a diverse range of models. Our findings demonstrate the promise of machine learning in combating scam profiles. Ensemble methods achieved superior performance across various evaluation metrics. This aligns with the observations made in related research, highlighting the effectiveness of ensemble methods for social media profile analysis.

This study contributes to the ongoing development of robust scam profile detection systems on Instagram in several ways. First, it underscores the effectiveness of ensemble methods for this specific task. Second, it emphasizes the importance of the dataset size, as our findings with a large dataset surpass the accuracy achieved in some studies using smaller datasets. Third, it highlights the need for the further exploration of feature engineering techniques to potentially enhance the model performance.

While this research offers valuable insights, several avenues exist for future exploration. First, incorporating techniques like k-fold cross-validation can offer a more resilient assessment of all models. Second, delving deeper into feature engineering by creating new features that capture the temporal evolution of profile characteristics or user behavior patterns could potentially improve the model performance. Third, exploring the integration of visual features alongside textual data, as suggested by Bharne and Bhaladhare [13], holds promise in potentially achieving even higher accuracy. Fourth, exploring the efficacy of advanced neural network architectures like CNNs for the analysis of images could be another avenue for future research, particularly when combined with textual data analysis. Finally, deploying the most effective models in a real-world setting and evaluating their performance on live Instagram data could offer valuable insights for practical implementation. By embarking on these prospective avenues of research, we can contribute to the development of increasingly accurate and reliable scam profile detection systems, fostering a safer and more trustworthy online environment for Instagram users.

In summary, this research marks a meaningful stride toward mitigating the threats posed by fake profiles on social media. By harnessing the capabilities of machine learning, we can enhance the digital ecosystem's resilience against fraudulent entities, thereby fostering a more secure and trustworthy online community.

Author Contributions: Conceptualization, B.G.B.; Methodology, B.G.B. and Q.L.; Formal analysis, B.G.B.; Investigation, B.G.B.; Writing—original draft, B.G.B.; Writing—review & editing, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: The research received no funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found at https://www.kaggle.com/datasets/krpurba/fakeauthentic-user-instagram (accessed on 4 March 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this article:

ML	Machine Learning
LR	Logistic Regression
DT	Decision Trees
SVM	Support Vector Machine
RF	Random Forest
KNN	K-Nearest Neighbors
XGBoost	eXtreme Gradient Boosting
GB	Gradient Boosting
AdaBoost	Adaptive Boosting
ET	Extra Trees
ROC	Receiver Operating Characteristic
F1	F1-Score
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
TPR	True Positive Rate
FPR	False Positive Rate
TN	True Negative
FN	False Negative

References

1. Adekunle, B.; Kajumba, C. Social media and Economic Development: The role of Instagram in developing countries. In *Advances in Theory and Practice of Emerging Markets*; Springer: Cham, Switzerland, 2021; pp. 85–99.
2. Akyon, F.C.; Kalfaoglu, M.E. Instagram Fake and Automated account Detection. In Proceedings of the 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 31 October–2 November 2019.
3. Crawford, M.; Khoshgoftaar, T.M.; Prusa, J.D.; Richter, A.; Najada, H.A. Survey of review spam detection using machine learning techniques. *J. Big Data* **2015**, *2*, 23. [CrossRef]
4. Dwivedi, Y.K.; Kelly, G.; Janssen, M.; Rana, N.P.; Slade, E.L.; Clement, M. Social Media: The Good, the Bad, and the Ugly. *Inf. Syst. Front.* **2018**, *20*, 419–423. . [CrossRef]
5. Sahoo, S.R.; Gupta, B.B. Fake profile detection in multimedia big data on online social networks. *Int. J. Inf. Comput. Secur.* **2020**, *12*, 303. [CrossRef]
6. Tiwari, V. Analysis and detection of fake profile over social network. In Proceedings of the 2017 International Conference on Computing, Communication and Automation, Greater Noida, India, 5–6 May 2017. Available online: https://ieeexplore.ieee.org/document/8229795 (accessed on 6 May 2020).
7. Subrahmanian, V.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; Menczer, F. The darpa twitter bot challenge. *Computer* **2016**, *49*, 38–46. [CrossRef]
8. Varol, O.; Ferrara, E.; Davis, C.A.; Menczer, F.; Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. *arXiv* **2017**, arXiv:1703.03107.
9. Zhang, X.; Zhu, S.; Liang, W. Detecting spam and promoting campaigns in the Twitter social network. In Proceedings of the Data Mining (ICDM), 2012 IEEE 12th International Conference, Brussels, Belgium, 10–13 December 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1194–1199.
10. Abinaya, R.; Naveen, P. Spam Detection On Social Media Platforms. In Proceedings of the 2020 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 23–24 July 2020. Available online: https://ieeexplore.ieee.org/document/9201948 (accessed on 1 December 2021).

11. Shafahi, M.; Kempers, L.; Afsarmanesh, H. Phishing through social bots on Twitter. In Proceedings of the Big Data (Big Data), 2016 IEEE International Conference, Washington, DC, USA, 5–8 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 3703–3712.
12. Boshmaf, Y.; Muslukhov, I.; Beznosov, K.; Ripeanu, M. The socialbot network: When bots socialize for fame and money. In Proceedings of the 27th Annual Computer Security Applications Conference, Orlando, FL, USA, 5–9 December 2011; pp. 93–102.
13. Bharne, S.; Bhaladhare, P. An Enhanced Scammer Detection Model for Online Social Network Frauds Using Machine Learning. *IJRITCC* **2023**, *11*, 239–249. [CrossRef]
14. Hakimi, A.N.; Ramli, S.; Wook, M.; Zainudin, N.M.; Hasbullah, N.A.; Wahab, N.A.; Razali, N.A.M. Identifying Fake Account in Facebook Using Machine Learning. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2019; pp. 441–450. [CrossRef]
15. Harris, P.; Gojal, J.; Chitra, R.; Anithra, S. Fake Instagram Profile Identification and Classification using Machine Learning. In Proceedings of the 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 1–3 October 2021. Available online: https://ieeexplore.ieee.org/document/9587858 (accessed on 3 February 2024).
16. Kadam, N.; Sharma, S.K. Social Media Fake Profile Detection Using Data Mining Technique. *J. Adv. Inf. Technol.* **2022**, *13*, 518–523. [CrossRef]
17. Shreya, K.; Kothapelly, A.; V, D.; Shanmugasundaram, H. Identification of Fake accounts in social media using machine learning. In Proceedings of the 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 26–27 December 2022. Available online: https://ieeexplore.ieee.org/document/10060194 (accessed on 8 September 2023).
18. Kaddoura, S.; Chandrasekaran, G.; Popescu, D.E.; Duraisamy, J.H. A systematic literature review on spam content detection and classification. *Peerj Comput. Sci.* **2022**, *8*, e830. [CrossRef] [PubMed]
19. Spam and Scams on the Rise: How to Detect Them and Protect Your Brand. Available online: https://blog.brandbastion.com/spam-and-scams-are-on-the-rise-how-to-detect-them-and-protect-your-brand (accessed on 21 March 2024).
20. Al-Zoubi, A.M.; Alqatawna, J.; Paris, H. Spam profile detection in social networks based on public features. In Proceedings of the 2017 8th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 4–6 April 2017. [CrossRef]
21. Purba, K.R.; Asirvatham, D.; Murugesan, R.K. Classification of instagram fake users using supervised machine learning algorithms. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 2763. [CrossRef]