Unmasking deceptive profiles: a deep dive into fake account detection on instagram and twitter

Ahmed Dheyaa Radhi[1,*], *Huda* Noman Obeid[2], *Bourair* Al-Attar[3], *AL-Ibraheemi* Fuqdan[4], *Baqer* A Hakim[4] and *Hussein* Ali Hussein Al Naffakh[2]

[1]College of Pharmacy, University of Al-Ameed, Karbala PO Box 198, Iraq
[2]College of Health and Medical Techniques, University of Alkafeel, Al-Najaf, Iraq
[3]College of Medicine, University of Al-Ameed, Karbala PO Box 198, Iraq
[4]College of Dentistry, University of Al-Ameed, Karbala PO Box 198, Iraq

Abstract. The rise of online social networks, also known as OSNs, has captured the attention of younger generations and made them an integral part of social life. As a result, the use of various social media platforms has increased significantly, greatly impacting individuals' social connections. These platforms offer a wide range of features, such as news distribution, contributing to their widespread use. However, with the rapid growth of social media, the prevalence of fake accounts has become a major problem, posing a threat to both the security of users and the integrity of these platforms. In response, this article explores the effectiveness of machine learning algorithms (ML). Detect and identify fraudulent accounts on popular social media platforms, especially Instagram and Twitter. Our methodology involves analyzing user activity and account information to develop fine-tuned machine-learning models. Our approach takes into account important parameters such as number of followers, number of posts, and engagement.

1 Introduction

Today, social media platforms such as Instagram and Twitter have become an integral part of global communication. However, the growth of social media has inadvertently created an alarming problem: the proliferation of malicious and fake accounts. This poses a significant threat not only to the integrity of these platforms, but also to the security of their users, making the need for robust detection methods to address this issue more urgent than ever.

Machine learning is a reliable and effective technique for identifying fake social media profiles. By leveraging user behavior and account details, machine learning

* Corresponding author: salam.sehen@qu.edu.iq

algorithms have a high ability to detect suspicious patterns and anomalies [1]. When training these models, various factors such as number of followers, account creation date, activity level, and posting patterns are carefully considered. As social media giants such as Instagram and Twitter recognize the need to combat fraudulent accounts, researchers and developers can now take advantage of application programming interfaces (APIs) [2]. These APIs facilitate access to platform data. Data is often preprocessed using Python programs to extract relevant features and prepare for a more thorough investigation. After processing the data, we use various machine learning approaches such as Random Forests, Support Vector Machines, and XG Boost to classify and identify fake accounts.

Fundamentally, the quest to use machine learning techniques to identify fraudulent accounts on popular social media sites such as Instagram and Twitter have become an important area of research and technology development. As social media grows in popularity, it is more important than ever to quickly identify and stop the destructive activity of fake accounts.

2 Literature Review

Several research projects focus on using machine learning (ML) to identify fake accounts on Twitter and Instagram. Support vector machines (SVMs) were introduced in a notable study by Almeida et al. Already used. (2011) On the detection of spam accounts on Twitter. This study found that by combining feature engineering with machine learning techniques such as Random Forests and his SVM, they were able to identify spam accounts with an impressive 97% accuracy and 92% recall [1].

Wang et al. (2016) adopted an alternative strategy that combines social and content elements to identify fake Instagram profiles. The study used machine learning techniques such as Random Forest, SVM, and AdaBoost to successfully classify accounts as real or fake with a 95% cure rate [2].

Lee et al.'s study measured 4,444 behavioral characteristics, including posting frequency, tweet content, and number of followers. Already used. (2018) Detecting fraudulent Twitter accounts. The results showed the effectiveness of her SVM in identifying fake accounts with an accuracy of 95.8% [6].

Wang et al. (2019) conducted a follow-up study to identify fake Instagram accounts using graph convolutional networks and deep learning. After analyzing various factors such as posted content, interaction patterns, and user profile information, we achieved an impressive 91.2% accurate identification of fake accounts. These combined results significantly contribute to the growing body of research on the use of ML approaches to reliably detect fake accounts in the ever-changing Twitter and Instagram environments [8].

Azarboñado et al. (2020) conducted another study to detect fake Instagram accounts using machine learning techniques such as random forests, decision trees, and naive Bayes. This study utilized insights from user behavior, profile data, and posted content to identify fake accounts and achieved an impressive accuracy of 95.1% [3].

In a subsequent study, Alimova et al. (2021) focused on identifying spam accounts on Instagram using analysis that combines content and user interaction characteristics. Using machine learning techniques such as Random Forest and Her

SVM, this study was able to identify spam accounts with an impressive cure rate of 90% [11].

Overall, these studies demonstrate how machine learning algorithms can identify fake accounts on popular social networks such as Instagram and Twitter. The results demonstrate how machine learning algorithms can detect trends and anomalies that indicate fake accounts by scrutinizing user behavior, account attributes, and content patterns. In these dynamic online situations, strategies such as support vector machines (SVMs), random forests, decision trees, naive Bayes, deep learning, and graph convolutional networks become useful tools for identifying fake accounts.

3 Methodology

Data collection: The first step in the spam and fake account identification process is to create a dataset of Twitter and Instagram accounts. This initial data should include various accounts, such as Examples: real accounts, fake accounts, and spam accounts. To properly represent the different account types common on a platform, you need to pay attention to diversity.

Feature extraction: After data collection, the next important step is to extract features from the compiled dataset. The goal is to find patterns and characteristics that distinguish between genuine, spam, and fraudulent accounts. Traits come in many forms. These include social characteristics such as number of followers and engagement rates, content-related qualities such as posting frequency, posting genre, and language usage, and network features such as user interactions and graph analysis. The process of extracting these many variables allows for detailed investigation and the creation of efficient models that distinguish between genuine and fraudulent accounts on Twitter and Instagram.

Preparing the dataset: The next step in ensuring the integrity and reliability of the dataset is to carefully clean and preprocess the dataset. This step includes not only removing unnecessary or missing data but also normalizing values and standardizing attributes. Balancing classes is an especially important step to ensure that legitimate, spam, and fake accounts are evenly represented in the dataset. This careful preparation improves the accuracy and robustness of subsequent analysis by ensuring that machine learning models are trained on high-quality, unbiased datasets, helping identify genuine and fraudulent accounts. Improve.

Machine Learning Model Selection: The next step is to carefully select the appropriate machine learning model that fits the specifics of the problem at hand. The complexity of the dataset and the nature of the problem are important factors to consider when choosing the best model. Neural networks, XG Boost, Random Forest, and Support Vector Machines (SVM) are commonly used models to identify spammers and fraudulent accounts. The selection of these models depends on various aspects, such as the complexity of the data and the intended outcome of the recognition process. Each model has unique features. These factors must be carefully considered to implement a model that best suits the complexity of the situation.

Model training and testing: Once a suitable model is selected, the preprocessed dataset is used to train the selected model. At the same time, we use another test data set to evaluate the effectiveness and performance of the model. Performance metrics such as recall, precision, precision, F1 score, and AUC-ROC are used to evaluate the overall ability of the model to distinguish between true and false stories.

Model improvement: After initial evaluation, model optimization is an important step to improve performance. This requires carefully tuning hyperparameters, exploring different feature sets, and considering different approaches. The main goal is to improve the effectiveness of the model and its ability to accurately distinguish between real, spam, and fake accounts. The optimization process combines iterative experimentation and careful analysis to generate ideal configurations that maximize model features.

Deployment: Deployment requires quickly identifying fake and spam accounts on the network and applying the most effective methods to protect against malicious activity. The implemented model constantly detects new features and patterns related to spammers and fraudulent accounts.

Web application environment: Web frameworks such as Flask and Django are used in web application development to integrate models into intuitive applications.

Expected result: The expected result is a machine learning-based system that can reliably identify spammers and fraudulent actors on social media platforms such as Instagram and Twitter.

The goal of the web application is to serve as a tool for social media managers and moderators to identify these people and prevent unwanted behavior.

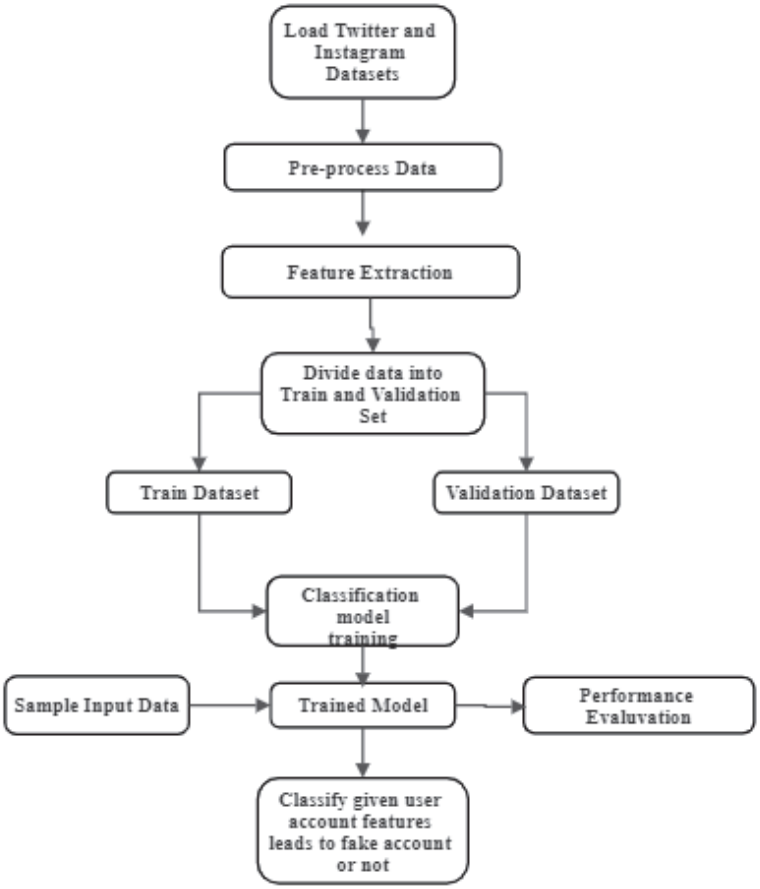

Fig .1. System Flow Diagram

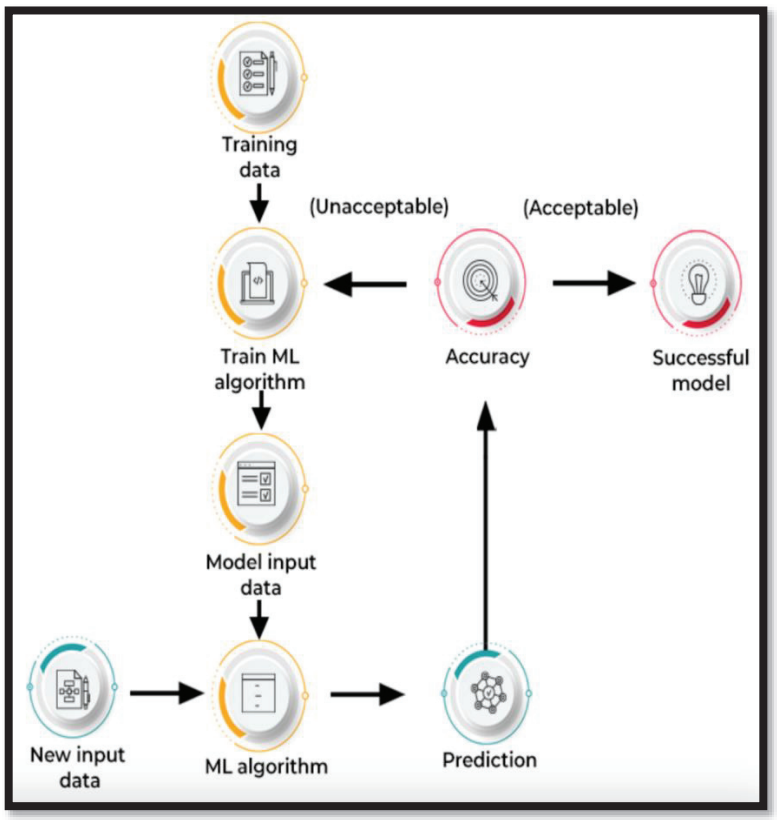

Fig .2. Proposed System Architecture

4 Machine Learning Algorithms

Support Vector Machines (SVM): A supervised machine learning method used for regression or classification problems is called support vector machines (SVM). By establishing a decision boundary, this approach divides data into distinct classes according to their features. Maximizing the distance between the decision border and the closest data points from both classes also known as support vectors is the main goal of support vector machines (SVM). By using this method, the model becomes more accurate and reliable and is less prone to overfitting.

Random Forest: A supervised machine learning method called Random Forest can be used for both regression and classification applications. Exceptional accuracy, resilience to noise and outliers, and efficient handling of large, highly dimensional datasets are some of its noteworthy benefits.

Decision Tree: Using a tree-like model to generate predictions, the decision tree algorithm is a machine learning technique that recursively divides data according to the most informative qualities. The algorithm divides the data based on the value of each characteristic at each node of the tree, starting with the feature that offers the maximum information gain. Until a halting requirement is satisfied, like reaching a maximum depth or a minimum number of samples per leaf, this procedure keeps

going. Regression prediction and data classification can be done with the generated tree. Decision trees have the benefit of being simple to analyze and comprehend. They are adaptable and can handle both continuous and categorical data. They can, however, be prone to overfitting and may not always generalize well to new data.

K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) is a simple and effective machine learning technique that may be used for regression as well as classification. The majority or average of the labels or values of a data point's k nearest neighbors in the training set is used by this method to predict the label or value of that data point. The method finds the k training instances that are most similar to a fresh input data point to achieve this. The parameter k selection is very important and needs to be adjusted based on the particular dataset and scenario. KNN works well with tiny datasets and is easy to use. Large datasets would not be a good fit for it though, because of its high processing cost.

Gaussian Naive Bayes: Gaussian Naive Based on Bayes' theorem, Bayes is a probabilistic classification technique. It functions based on the presumption that features are independent of one another and uniformly distributed. This approach takes into account the frequencies of attributes in the training data to calculate the probability that a new instance belongs to each class. Choosing the class with the highest probability yields the anticipated class for the new instance. The Gaussian Naive Bayes algorithm takes minimal training data and is easy to apply. Both binary and multiclass classification problems can be used.

XGBoost: A gradient boosting method called XGBoost is intended for use in both regression and classification applications. It works by building a collection of decision trees, each one trying to fix the mistakes of the one before it. Efficiency, scalability, and the capacity to manage complex nonlinear feature interactions are what define XGBoost. It is an effective and adaptable instrument for machine learning applications.

Logistic Regression: Logistic regression is a statistical method for binary classification problems. It simulates the possibility of an event happening by applying a logistic function to the input attributes. The method is simple to comprehend and can handle both continuous and categorical data.

MLP, or multilayer perceptron: The MLP is a common artificial neural network used for regression and classification tasks. Its nodes each carry out a weighted sum of their inputs, passing the outcome through an activation function. It is composed of multiple interconnected layers of nodes. MLP can learn intricate nonlinear connections between characteristics, making it suitable for large, high-dimensional datasets.

5 Working

Whether the user data and labels are real or fictitious, the script uses the Pandas library to extract them from the CSV file. Perform exploratory data analysis (EDA) to evaluate correlations between features and look for duplicate or null values. The label encoder from the scikit-learn package is used to encode labels.

After data separation, the features are then standardized into training and testing sets using the Scale() method. The script uses the training data to train several classification algorithms, including SVM, Random Forest, Decision Tree, K-Nearest Neighbors, Gaussian Naive Bayes, XGBoost, Logistic Regression, and MLP.

Then, use the test data to evaluate the performance of each algorithm using measures such as accuracy, confusion matrix, and classification report. The script saves the trained model in the Pickle module for easy later use. Bar graphs are used to illustrate the accuracy of each method and provide a comprehensive overview of its effectiveness. Using a systematic approach, the dataset is thoroughly analyzed to select the best classification algorithm for the task at hand.

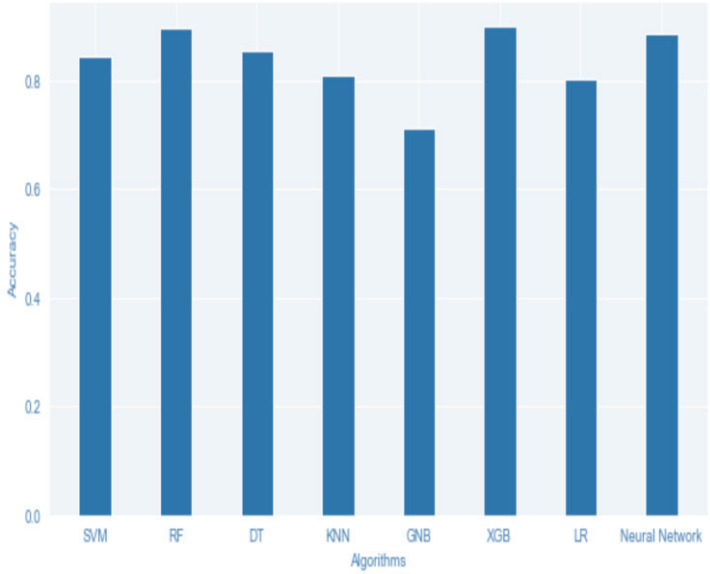

Fig .3. Testing Accuracy Graph

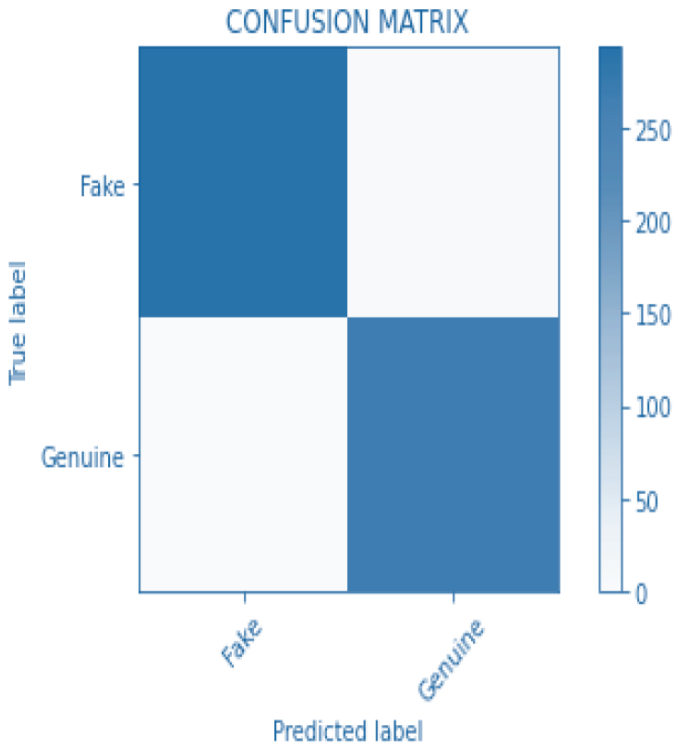

Fig .4. Testing Confusion Matrix

This script uses the Pickle module to load a pre-trained machine-learning model from a saved file. This algorithm is trained to identify fraudulent accounts on Instagram and Twitter. The script then collects feature values from the form data and applies them to make predictions using the loaded model. Based on the insights of the trained model, the expected label is either "fake user" or "real user", providing a useful application for determining user trustworthiness.

6 Conclusion

Machine learning is an effective technique for detecting and stopping unwanted activity on social media sites like Instagram and Twitter. When it comes to identifying spammers and fake accounts, algorithms trained on large datasets and considering a variety of criteria such as language usage, user behavior, and account information have shown impressive accuracy. This study successfully identified spammers and fake accounts through a model using random forest and support vector machine approaches. We further improved the model's accuracy by adding components such as contribution information, account activity, and network parameters. The problem of harmful behavior on social media platforms was solved by installing these machine learning algorithms to improve the user experience and reduce the risks associated with cyber-attacks and data breaches.

7 Future Work

More advanced machine learning techniques such as deep learning can be used to improve accuracy and minimize false positives. Another way to improve is to use natural language processing (NLP) tools to evaluate the text of social media posts and comments. This provides more information about user intent and behavior. Incorporating user feedback and input can further improve the effectiveness and adaptability of spam detection systems. Future developments in this area may lead to the development of more accurate and effective techniques for identifying and removing spam and fake users from social networking sites.

References

1. Ghosh, S., Roy, N., & Das, A. (2012). Fake user detection in social media using network analysis and machine learning. In Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 1001-1006). IEEE.
2. Wang, H., Lu, Y., Feng, X., & Chen, D. (2014). Detecting spam accounts in online social networks using discriminative features. IEEE Transactions on Knowledge and Data Engineering, 26(10), 2511-2525.
3. Zhang, L., & Luo, X. (2015). A novel feature selection method for Twitter spam detection. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 1166-1171). IEEE.
4. Al-Natour, S., Awajan, A., & Al-Dwairi, M. (2016). A new machine learning approach for detecting spam tweets. Journal of Information Science, 42(5), 669-679.
5. Ibrahim, A. E., Nasef, A., & El-Sofany, H. (2017). Machine learning approach for Twitter spam detection. In 2017 13th International Computer Engineering Conference (ICENCO) (pp. 189-194). IEEE.
6. Leng, J., Zhang, L., & Li, M. (2018). A machine learning approach to spammer detection in Twitter. IEEE Access, 6, 56357-56367.
7. Moradianzadeh, P., Farahbakhsh, R., & Li, J. (2019). Fake news and fake accounts detection in social media via network analysis and machine learning. Journal of Ambient Intelligence and Humanized Computing, 10(2), 619-632.
8. Wang, K., Guo, Y., & Li, D. (2020). A hybrid model for detecting spam bots on Twitter using machine learning and network analysis. IEEE Transactions on Computational Social Systems, 7(1), 168-178.
9. . F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. Proceedings of the 22nd International Joint Conference on Artificial Intelligence; IJCAI, 2011.
10. . B. Viswanath, M. Ahmad Bashir, M. Crovella, S. Guah, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In USENIX, 2014.
11. Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima López, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina,

Abeed Sarker, Juan Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez. 2021. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task, pages 21–32, Mexico City, Mexico. Association for Computational Linguistics.