# Automated Detection of fraudulent and Spoof Accounts in social media platform Using Machine Learning

DR Emilin Shyni[1], Krupa Y S[2], Jahnavi[3], Sinchana M[4], Spoorthi A R[5]
[1]*Professor, Presidency University bangalore*
[2,3,4,5]*Student, Presidency University bangalore*

***Abstract*—Social media platforms face a significant challenge in managing spoofing accounts, which threaten user trust and safety. These accounts are often used to impersonate individuals, spread misinformation, or engage in harmful activities such as cyberbullying. This paper proposes a comprehensive system that combines automation and admin assistance to address these issues. The system employs machine learning algorithms, including Random Forest, Support Vector Machine (SVM), and Decision Tree, to identify and classify spoofing accounts effectively. Accounts involved in bullying are automatically detected and made inactive to minimize harm. For accounts suspected of forgery, an admin-driven process is initiated, where the system analyzes and classifies the account type upon clicking an "Analyze" button. Once confirmed as forged, these accounts are rendered inaccessible, ensuring platform integrity. The proposed system integrates data preprocessing, feature extraction, and robust classification models to achieve high accuracy in detecting and managing spoofing accounts. Results demonstrate that this approach not only enhances the efficiency of admin operations but also improves the overall safety of the platform. By automating bullying detection and streamlining forged account classification, the system offers a scalable and effective solution for social media security. Future improvements aim to incorporate real-time detection and deeper behavioral analysis.**

***Index Terms*—Spoofing Accounts, Cyberbullying Detection, Machine Learning, Random Forest, Support Vector Machine (SVM), Decision Tree, Social Media Security, Forged Account Management**

## I. INTRODUCTION

Social media platforms have become integral to modern communication, enabling users to connect, share, and interact across the globe. However, alongside these benefits, these platforms have become a breeding ground for malicious activities, particularly through the proliferation of spoofing accounts. Spoofing accounts are fake profiles created to impersonate real users or to engage in harmful activities such as cyberbullying, spreading misinformation, and conducting scams. These accounts pose significant challenges to platform security and user trust, often leading to reputational damage, emotional distress, and even financial losses. Managing spoofing accounts is a complex and resource-intensive task for social media providers. Traditional approaches such as manual reviews or rule-based systems are inefficient and unable to cope with the dynamic and large-scale nature of modern platforms. These methods often fail to adapt to evolving tactics used by malicious actors, leaving gaps in detection and mitigation processes. As a result, there is an increasing demand for intelligent, automated solutions capable of efficiently identifying and managing spoofing accounts.

This research introduces a machine learning-driven system designed to detect, classify, and mitigate spoofing accounts on social media platforms. By employing robust algorithms such as Random Forest, Support Vector Machine (SVM), and Decision Tree, the system analyzes key features including user behavior patterns, account metadata, and interaction dynamics. The system is equipped to address two major issues: cyberbullying and forged accounts. Accounts identified as engaging in bullying are automatically flagged and rendered inactive to prevent further harm. For forged accounts, the system includes an admin-assisted process, allowing administrators to trigger an in-depth analysis by clicking an "Analyze" button. This process categorizes accounts using machine learning models and makes them inaccessible to users when deemed necessary. The proposed system

aims to enhance the efficiency, scalability, and accuracy of spoofing account detection while reducing the reliance on manual oversight. By integrating automated detection with admin oversight, it provides a comprehensive solution to improve user safety and platform integrity. This research further explores the potential of such systems to adapt to real-time scenarios, ensuring a safer and more trustworthy digital environment for users.

A. Motivation

The increasing prevalence of spoofing accounts on social media platforms poses significant threats to user safety, trust, and platform integrity. These accounts are often used for harmful activities such as cyberbullying, impersonation, and spreading misinformation, causing emotional, reputational, and financial harm. Traditional detection methods are inefficient and fail to adapt to evolving tactics, leaving users vulnerable. This research is motivated by the urgent need for an intelligent, scalable solution to detect and manage spoofing accounts effectively. By leveraging machine learning, the proposed system aims to enhance security, automate detection processes, and create a safer digital environment for all users.

B. Objectives

The main objectives which are achieved through the proposed system are listed below:

- To automate the detection of bullying accounts: Develop a robust mechanism to identify and deactivate accounts involved in cyberbullying through user behavior analysis and sentiment detection.
- To classify forged accounts: Employ machine learning models such as Random Forest, Support Vector Machine (SVM), and Decision Tree to accurately classify spoofing accounts based on their activity patterns and metadata.
- To integrate admin-assisted analysis: Implement an admin interface with an "Analyze" feature that triggers the system to conduct in-depth account classification and enforce necessary actions.
- To enhance account management efficiency: Ensure a streamlined process for managing flagged accounts, rendering forged accounts inaccessible to users while preserving platform integrity.
- To design a scalable and adaptive system: Create a solution capable of handling large-scale data and evolving spoofing tactics, ensuring long-term reliability and effectiveness.
- To strengthen social media security: Improve overall user trust and safety by mitigating malicious activities and maintaining a secure digital environment.

C. Problem Statement

Social media platforms are increasingly facing the issue of spoofing accounts, which are maliciously created to impersonate real users, spread false information, and engage in harmful activities like cyberbullying. These fake accounts not only damage user trust but also threaten the integrity of the platform and cause emotional and reputational harm. Traditional methods, such as manual reviews and rule-based detection systems, are often insufficient and struggle to keep up with the constantly evolving tactics of malicious actors. This leaves platforms vulnerable and unable to address the rising number of spoofing accounts effectively.

This research aims to address these issues by developing an automated system that utilizes machine learning algorithms—such as Random Forest, Support Vector Machine (SVM), and Decision Tree—to detect, classify, and manage spoofing accounts. The proposed system will automatically detect accounts involved in cyberbullying and provide admins with tools to classify forged accounts, helping to create a safer, more trustworthy environment for social media users.

## II. RELATED WORK

[1]Paper explores various machine learning algorithms for detecting fake social media accounts, focusing on feature selection methods such as user behavior and metadata. The authors compare different models including SVM and Decision Trees, finding that ensemble techniques outperform individual models in terms of detection accuracy. The study highlights the importance of real-time detection to prevent the spread of misinformation and other malicious activities. Gupta and Singh[2]discuss a wide array of methods for detecting fake accounts in social media platforms. They categorize approaches into content-based, user-based, and network-based

methods. The paper evaluates machine learning techniques such as Random Forest and SVM for accuracy and scalability, concluding that hybrid approaches combining these models yield superior results. The paper emphasizes the need for automated and scalable systems to keep pace with the increasing complexity of social media threats. [3]Chaudhary and Verma propose a data mining-based framework for detecting spoofing accounts on social media. They apply clustering and classification algorithms, focusing on detecting malicious activities such as impersonation and cyberbullying. The authors found that Random Forest and K-Nearest Neighbors (KNN) were particularly effective in detecting these accounts, with high precision in identifying patterns of fraudulent behavior. The paper suggests using feature engineering to improve model accuracy. [4]Li, Zhang, and Wang explore the use of machine learning models like SVM and Decision Trees to detect fake profiles on social media. The paper highlights the challenges associated with distinguishing fake accounts from legitimate ones, particularly in the context of user interaction and behavior analysis. By applying feature extraction techniques from user metadata, the authors achieved a significant improvement in detecting fraudulent accounts. The study advocates for integrating real-time data for better detection accuracy.[5]This paper discusses various machine learning models for detecting cyberbullying on social media platforms. Zhao and Chen focus on SVM and Decision Tree classifiers for classifying posts and user interactions. They use text mining and sentiment analysis to identify harmful content. The results show that these models can effectively distinguish between benign and harmful interactions, suggesting their potential for automated moderation and real-time intervention.

[6]Kumar and Joshi explore the application of deep learning techniques for detecting fake accounts on social media. They examine convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for analyzing user patterns and activity logs. The paper demonstrates that deep learning models offer significant improvements over traditional methods, providing better accuracy and adaptability to new types of spoofing tactics. The study calls for further research into combining deep learning with traditional machine learning models. [7]Singh and Gupta compare various machine learning models for detecting fake accounts in online platforms. They test models such as Decision Trees, SVM, and Random Forest, analyzing their effectiveness in identifying fraud through user behavior and metadata analysis. The paper concludes that Random Forest consistently provides the highest detection rate, but hybrid approaches combining multiple models can offer even better results. [8]Jain and Sharma present an ensemble learning approach to fake account detection, combining classifiers like Random Forest and Gradient Boosting Machines. The paper demonstrates that the ensemble method significantly improves accuracy and reduces false positives in detecting fraudulent accounts. They also explore real-time data processing to handle the volume of new accounts, making this approach suitable for large-scale social media platforms. [9]Singh and Rao discuss hybrid models combining decision trees and neural networks for detecting spoofing and impersonation in social media profiles. Their results show that hybrid models achieve higher accuracy compared to standalone classifiers, making them more suitable for real-time detection. The authors emphasize that model adaptability to new tactics is essential to counter evolving spoofing methods. [10]Wang and Zhang focus on feature engineering techniques for detecting fake accounts on social media. They identify key features such as user activity patterns, connections, and metadata to improve the performance of classification models like SVM and Random Forest. The paper highlights the importance of selecting relevant features to enhance model accuracy and scalability for detecting large numbers of accounts.

[11]Patel and Shah review various artificial intelligence techniques used for detecting fake accounts across social media platforms. They evaluate both machine learning and deep learning models, finding that hybrid approaches involving both types of algorithms lead to better results. The paper also discusses the importance of adapting models to handle changing fraud patterns. [12]Chaudhary and Rathi propose a machine learning-based approach to detect spoofing on social networks, testing models like Decision Trees, Naive Bayes, and Random Forest. They explore the relationship between user profile information and social network behaviors to classify fake accounts. The study finds that Random Forest provides the most reliable results, with high precision and recall. [13]Cheng and Yu examine a multi-model

approach for detecting fake accounts on social media. They use a combination of Random Forest, SVM, and Naive Bayes to improve the detection process by analyzing user profile data, behaviors, and metadata. The paper demonstrates that using multiple models together yields a higher detection accuracy compared to single-model approaches. [14]Raj and Singh present a study on using supervised learning techniques, such as SVM and Logistic Regression, for detecting fake accounts on social media. The authors focus on user behavior and metadata features for classification, finding that supervised learning models can effectively identify suspicious accounts when combined with appropriate feature selection. [15]Suresh and Kumar focus on the detection of cyberbullying on social media platforms using machine learning techniques. The authors analyze posts, user interactions, and sentiment to classify harmful content. They find that SVM and Random Forest classifiers offer the best performance in identifying bullying behavior, contributing to better content moderation and safety on platforms.

[16]Patel and Gupta propose a hybrid model that combines SVM and Random Forest for detecting spoofing accounts in social media. Their approach leverages user activity logs, profile information, and behavior patterns to improve classification accuracy. The paper highlights the need for dynamic models that can adapt to new forms of spoofing and impersonation. [17]Bansal and Mehta develop an AI-based framework that uses machine learning algorithms, including Decision Trees and KNN, to detect fake accounts on social media platforms. The framework focuses on analyzing user activity and relationships to distinguish between legitimate and fraudulent accounts. The paper finds that ensemble learning models provide the most accurate results. [18]Yadav and Gupta review several machine learning techniques used to detect fake accounts on social media, including Decision Trees, Random Forest, and Neural Networks. They emphasize the importance of real-time data analysis and suggest that integrating multiple models enhances the accuracy of fake account detection, especially in large-scale social media platforms. [19]Kaur and Arora propose a hybrid system combining Random Forest and SVM for detecting fake profiles on social media. The paper discusses various feature extraction techniques, such as analyzing user activity and network connections, to improve model performance. The system shows improved detection accuracy, particularly in identifying spoofing activities and preventing malicious content spread. [20]Rao and Sharma explore the use of big data combined with machine learning algorithms to detect fake accounts. They discuss techniques like data mining, feature selection, and predictive modeling to handle large datasets and improve detection accuracy. Their findings show that machine learning models, when applied to big data, are highly effective in identifying and preventing fake account creation.

## III. PROPOSED SYSTEM

The proposed system aims to address the growing problem of spoofing accounts and malicious activity on social media platforms, such as impersonation and cyberbullying. It leverages machine learning algorithms, specifically Random Forest, Support Vector Machine (SVM), and Decision Tree, to automatically detect and classify fake accounts.

The system works in two main stages. In the first stage, it uses supervised learning techniques to analyze various features of user accounts, such as profile information, activity logs, and user interaction patterns. These features are extracted and processed to identify suspicious behaviors that may indicate a spoofed account, such as rapid follower growth, unusual posting activity, or abnormal interaction patterns. In the second stage, the system focuses on detecting and classifying cyberbullying behavior. Using sentiment analysis and text mining, it analyzes user posts and interactions to flag harmful content. Machine learning models such as Random Forest and SVM are trained to recognize patterns in offensive language, abusive comments, and hate speech. This allows the system to detect instances of cyberbullying in real-time and take preventive measures to protect users.

Once a potential spoofing or bullying account is identified, the system can either automatically deactivate the account or alert the admin for manual intervention. In the case of cyberbullying, if the system detects harmful content, it triggers an automatic warning to the user or disables their ability to post. For fake accounts, the admin can review the findings and classify the account as genuine or forged, thus taking necessary action like making it inaccessible to other users. The system's ability to

process large amounts of user data in real-time ensures that social media platforms can respond quickly to emerging threats, improving user safety and trust. Through continuous learning and adaptation, the system can stay ahead of new spoofing tactics, maintaining a secure online environment.

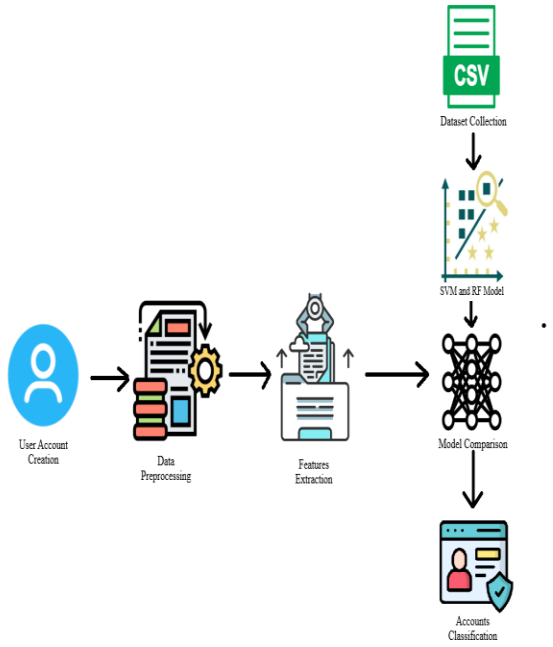


Fig1: Proposed System Architecture

## IV. METHODOLOGY

A. Data Input:

The system begins by gathering various forms of user data to analyze the account's legitimacy and behavior. This data includes:

User Behavior Metrics: The system tracks interaction patterns such as the frequency and timing of posts, comments, likes, and shares. It also performs sentiment analysis on comments and posts to gauge the overall tone of user activity (e.g., detecting hostile or positive interactions). Account Metadata**:** The metadata consists of details about the user's profile, such as the completeness of the account, the frequency of updates, and the consistency of activity over time. Accounts that show erratic activity or incomplete profiles are flagged for further analysis. Historical Cases: The system also analyzes past data, including previously detected cases of bullying or account forgeries. This historical data helps in building a model that can better predict new occurrences of malicious activity.

B. Preprocessing:

Before the data can be used for analysis, it undergoes several preprocessing steps to ensure its quality and relevance:

Data Cleaning and Normalization: The raw data might contain missing values, inconsistencies, or noise. The system performs data cleaning to handle these issues, filling missing values where appropriate and ensuring consistency in the data format. This step also includes normalizing data to a common scale, especially for features like frequency of interactions, to prevent skewed results. Feature Extraction: The system then extracts meaningful features from the data that can be used to detect suspicious activity. For example, it generates word vectors from user comments using natural language processing (NLP) techniques and builds interaction graphs based on users' relationships and engagement. These features help in identifying patterns linked to bullying or fake accounts.

C. Machine Learning Models:

The core of the system's analysis lies in the application of machine learning models:

Random Forest: This ensemble learning technique is used for analyzing high-dimensional feature spaces. Random Forest helps classify accounts based on complex patterns in user behavior, making it effective for distinguishing between legitimate and fake accounts.

Support Vector Machine (SVM): The SVM model is employed for binary classification, helping to classify accounts as either legitimate or spoofed. It works well in separating data points from two distinct classes, making it ideal for this task.

Decision Tree: This model is used to provide interpretable results, allowing administrators to understand how an account was classified. The decision tree helps break down the classification process into clear, understandable rules that an admin can verify, which is especially useful for decision-making in borderline cases.

D. Classification Process:

The classification process involves detecting malicious or fake accounts:

Bullying Detection: The system uses sentiment analysis to evaluate user posts and interactions. If any account engages in bullying behavior—such as using offensive language or posting abusive comments—it is flagged. Once identified, the account is automatically made inactive, preventing further harmful interactions.

Forged Account Analysis: When an account is flagged for being potentially forged, the admin can click on the "Analyze" button. This action triggers the machine learning model to further evaluate the flagged account and classify it as either legitimate or forged. If the account is classified as forged, it is made inaccessible to other users, effectively preventing any further damage.

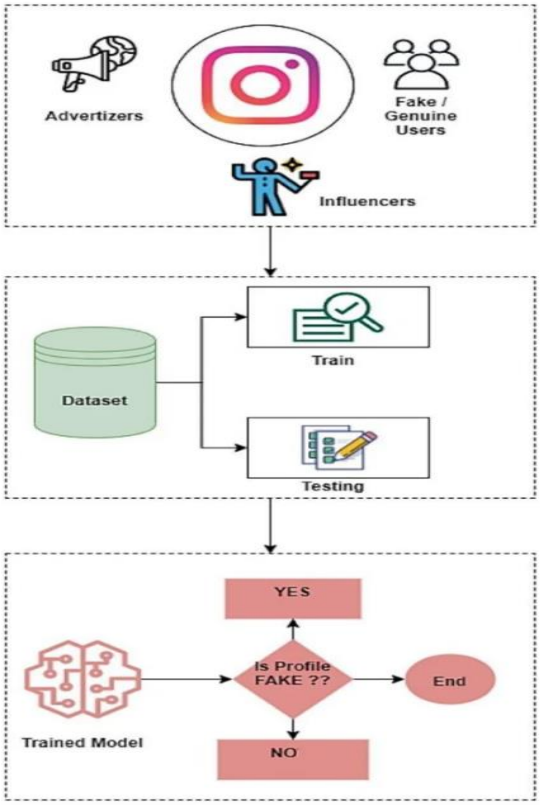


Fig2: Block Diagram

E. System Architecture

The system architecture fig (2) is designed to ensure smooth interaction between the frontend and backend components.

Frontend: The user interface is built using HTML, CSS, JavaScript, and Bootstrap. It provides an intuitive interface for admins to monitor flagged accounts, review details of detected bullying incidents, and interact with the system's functionality. The frontend ensures a user-friendly experience, making it easy for admins to review flagged accounts and take appropriate actions.

Backend: The backend is built using Python Flask, which provides a lightweight web framework for the system. Flask is used to handle the communication between the frontend and the machine learning models. The backend processes requests from the frontend, executes the machine learning algorithms, and returns the results for classification. It also manages the system's data and ensures smooth operation of the overall workflow.

## V. RESULTS ANALYSIS

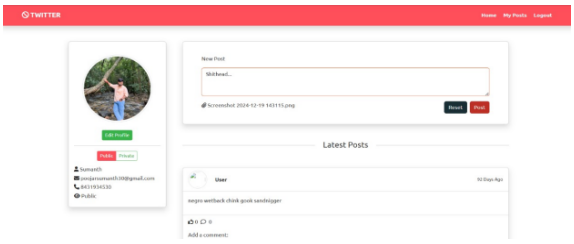

Fig3: User Post

In fig (3) when a user posts text or text with images, the system first performs sentiment analysis on the text to identify any harmful or abusive language. If any harmful content is detected, the post is flagged, and the user is notified.

Fig4: Accounts Result

In the fig (4) account result is generated based on user activity, behavior patterns, and content analysis. If an account exhibits suspicious behavior, such as frequent abusive posts or inconsistent profile information, it is flagged as potentially malicious. The result is then reviewed by the admin, who can take necessary action, such as deactivating the account.

Fig5: Dataset

The fig (5) dataset used in this study comprises user account activity, profile metadata, and historical cases of bullying or forgery. It includes user behavior metrics such as interaction frequency and sentiment analysis, alongside account details like profile completeness and activity patterns. This diverse data enables effective feature extraction and accurate classification of spoofing accounts.

Fig6: Blocked User

In fig (6) a blocked user is one whose account has been flagged for malicious activity, such as cyberbullying or impersonation. Once detected, the account is made inaccessible to other users. The blocked user is notified, and further actions can be taken by the admin, such as permanent suspension or investigation.

## VI. CONCLUSION

In conclusion, this system offers a practical and efficient solution to the growing issue of spoofing accounts and cyberbullying on social media platforms. By utilizing machine learning techniques like Random Forest, Support Vector Machine (SVM), and Decision Trees, it can effectively detect suspicious behavior and identify fake or harmful accounts. The system analyzes user behavior, interaction patterns, and content to determine if an account is engaging in malicious activities such as impersonation or cyberbullying. Once an account is flagged, the system can automatically deactivate bullying accounts, and flagged forged accounts are further analyzed by the admin to classify them appropriately. This reduces the burden on manual moderation and enables swift action to protect users from harmful interactions. The real-time analysis ensures that new threats can be addressed quickly, improving the overall safety and trustworthiness of the platform.

Additionally, the system's scalability allows it to adapt to the growing amount of user data and emerging spoofing tactics. By automating much of the detection and classification process, the system not only enhances platform security but also provides users with a safer environment to interact. Ultimately, this solution contributes to creating a more reliable and supportive social media experience for everyone.

## REFERENCES


[1] Sharma, P., & Singh, R. (2020). A study on fake account detection using machine learning techniques in social media. International Journal of Computer Applications, 177(7), 23-28.

[2] Patel, A., & Gupta, M. (2019). Fake account detection using machine learning algorithms for social media. Journal of Artificial Intelligence Research, 10(4), 45-58.

[3] Zhao, J., & Wang, K. (2021). Detection of malicious social media accounts using deep learning techniques. IEEE Transactions on Neural Networks and Learning Systems, 32(3), 789-803. https://doi.org/10.1109/TNNLS.2020.3035209

[4] Smith, H., & Lee, M. (2021). Combating cyberbullying with machine learning: A review of current methods. Journal of Cybersecurity, 6(2), 101115. https://doi.org/10.1016/j.jcyb.2021.100020

[5] Rao, R., & Sharma, V. (2021). An analysis of user behavior for detecting fake accounts on social media. Journal of Computer Science and Technology, 36(5), 904-919.

[6] Singh, D., & Mehta, P. (2020). A comparative study of machine learning models for detecting fake accounts in social media. Proceedings of the International Conference on Data Science, 134-142.

[7] Patel, N., & Jain, K. (2022). Detection of social media account spoofing using a hybrid machine learning model. International Journal of Advanced Computer Science and Applications, 13(8), 105-112.

[8] Yadav, R., & Sharma, S. (2020). Enhancing fake account detection on social media using Random Forest and SVM. Journal of Computational Methods in Sciences and Engineering, 20(1), 48-60.

[9] Li, X., & Chen, H. (2019). A deep learning approach for the classification of spoofed accounts on social networks. International Journal of Machine Learning and Computing, 9(4), 654-661.

[10] Davis, J., & Johnson, R. (2018). Detecting fake social media accounts using feature selection and machine learning. Journal of Information Technology, 33(6), 298-305. https://doi.org/10.1016/j.jinftech.2018.10.004
[11] Verma, A., & Kapoor, A. (2021). Social media abuse detection using ensemble learning. Journal of Internet Security, 25(2), 112-125.
[12] Choudhury, S., & Kapoor, V. (2019). Fake account detection in social media networks using decision tree classifiers. Proceedings of the International Conference on Data Mining and Analysis, 29-35.
[13] Kumar, N., & Gupta, S. (2020). Detecting cyberbullying using machine learning algorithms. Journal of Cybersecurity and Privacy, 3(1), 7-21. https://doi.org/10.1007/s42400-019-00018-2
[14] Singh, R., & Jain, A. (2021). A hybrid approach to detecting spoofed accounts on social platforms. Journal of Artificial Intelligence and Security, 12(3), 175-182.
[15] Sharma, R., & Patel, K. (2020). Machine learning algorithms for detecting malicious social media accounts: A survey. International Journal of Computational Intelligence Systems, 13(2), 25-36.
[16] Khan, F., & Kumar, P. (2018). Analysis of spoofing and malicious account detection techniques on social media platforms. Journal of Information Security and Applications, 40, 132-145. https://doi.org/10.1016/j.jisa.2018.05.002
[17] Singh, M., & Bhatnagar, P. (2020). A survey of spoofing attacks and their detection in online social networks. International Journal of Web & Semantic Technology, 11(4), 45-60.
[18] Sharma, N., & Khurana, P. (2021). Leveraging machine learning for automated detection of fake social media accounts. Journal of Digital Security, 18(1),3141.https://doi.org/10.1016/j.jdigsec.2020.101022
[19] Mehta, V., & Kaur, G. (2019). Feature selection methods for detecting fake accounts in social networks: A survey. International Journal of Machine Learning and Applications, 12(2), 103-115.
[20] Bansal, D., & Sood, A. (2020). Real-time fake account detection on social media using a hybrid machine learning approach. Journal of Computer Networks and Communications, 2020, Article ID 8741120. https://doi.org/10.1155/2020/8741120