RESEARCH ARTICLE OPEN ACCESS

A Machine Learning-Based Approach for Spam Detection and Fake Account Identification on Social Media Platforms

[1]Vivek Kumar Jethani, [2]Dr. Vibhakar Pathak, [3]Dr. Vishal Shrivastava

[1]Department of CSE, Arya College of Engineering and IT, Jaipur, Rajasthan, India.

[2]Department of IT, Arya College of Engineering and IT, Jaipur, Rajasthan, India

[3]Department of CSE, Arya College of Engineering and IT, Jaipur, Rajasthan, India,

Abstract: With the exponential growth of social media and online communication platforms, spam messages and fake accounts have become a major challenge, affecting user experience, privacy, and digital security. This research presents an effective machine learning-driven system for identifying spam comments on YouTube and detecting fake accounts on Facebook. Various supervised learning algorithms, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naïve Bayes, are employed to analyze user activity, text content, and account metadata. A user-friendly application is also developed to allow real-time verification of comments and accounts. The system addresses key challenges such as evolving spam tactics, false positives, and high-volume data processing using scalable and adaptive machine learning models. Experimental results demonstrate that the proposed system achieves high accuracy in classifying spam and fake accounts while minimizing false positives. This solution contributes to building safer and more trustworthy online platforms by efficiently identifying malicious activities and enhancing the overall quality of user interactions.

Keywords: Machine Learning, Artificial Intelligence, Spam Detection, Fake Account Detection

1. Introduction

With the rapid expansion of online platforms, maintaining the authenticity of user interactions has become more critical than ever. Social media sites like YouTube and Facebook frequently face challenges from spammers and fake users who spread false information, unwanted advertisements, and harmful content. Fake accounts are also created to boost engagement artificially, promote scams, or carry out cybercrimes. Such activities not only interfere with real user experiences but also threaten the security of both users and platforms. Therefore, detecting and stopping these actions is essential to create a secure and reliable online space.

In the past, platforms primarily relied on rule-based techniques to detect spam and fake accounts. These approaches used fixed rules, like spotting certain keywords, blocking specific users, or monitoring repeated actions. However, they often fall short because spammers and fraudsters continuously adapt their tactics to bypass detection. To overcome these challenges, machine learning algorithms have become a more reliable and an efficient method for identifying spams and fake accounts.

Machine learning models are capable of processing large datasets and uncovering hidden patterns that distinguish real users from fraudulent ones. Popular algorithms used for this purpose include Logistic Regression, Random Forest, Support Vector

Machine (SVM), Naïve Bayes, among others.

These models continuously learn from historical data, allowing them to become more effective at spotting new forms of spam and fake accounts over time. By analyzing elements like text content, user activity, and account information, machine learning techniques can reliably identify whether a comment is spam or if a Facebook account is genuine or fake.

This research aims to develop a machine learning-driven system for detecting spam comments on YouTube and spotting fake accounts on Facebook. A user-friendly application has also been created to help users easily verify comments and accounts. Through experiments, various machine learning models were evaluated, demonstrating high accuracy in identifying fraudulent activities.

2. Spam Detection and Classification

The expansion of the internet and online platforms has brought many advantages, but it has also caused a significant rise in unwanted, harmful, or irrelevant content known as spam. Spam poses a serious problem across multiple online services, including email, social media, and websites. It typically involves sending bulk messages or content, often with malicious intent or to promote unrelated products and services. Tackling spam is essential to ensure a safe, smooth, and positive online experience.

Spam generally refers to any unsolicited or irrelevant content sent to a large audience, usually aimed at advertising products, promoting services, or spreading false information. Although it is commonly linked to email, spam also appears frequently on social media networks, messaging platforms, comment sections, and even search engine results. Spam can take many forms, including:

- **Advertising and Promotions:** Spam often includes promotional messages that users did not request, such as advertisements for products, services, or websites.
- **Phishing Links:** These are fraudulent messages that attempt to trick users into revealing sensitive information, such as passwords, credit card numbers, or personal details. Phishing scams can be extremely dangerous and lead to identity theft or financial loss.
- **Malware and Viruses:** Some spam messages contain harmful attachments or links that, when clicked, infect the user’s device with malware or viruses. These can lead to data theft, system damage, or loss of personal information.
- **Scams and Fraud:** Spam is frequently used to promote fraudulent schemes, such as "get-rich-quick" offers, fake job opportunities, or lottery scams. These messages often aim to deceive users into paying money or providing personal data.
- **Misinformation and Hoaxes:** Spam can also be used to spread false information, rumors, or conspiracy theories. These can cause confusion, panic, and damage public trust, especially in areas like health, politics, and safety.

3. Impact of Spam

Spam causes many problems for both users and online platforms. It not only makes it difficult for users to find useful content but also creates security risks and increases costs for service providers.

- **Poor User Experience:** Spam clutters online spaces, making it hard for users to find relevant content. In emails, spam messages take up space, causing important emails to be missed. On social media, spam comments disrupt discussions, making conversations less meaningful. Websites and forums filled with spam lose quality, making users less likely to engage.

- **Security Risks:** Spam can be dangerous, especially when used for phishing. Attackers send fake emails or messages to trick users into giving away personal information like passwords or bank details. Some spam messages contain malware, which can infect devices and steal data. Scammers also use spam to commit financial fraud by convincing users to send money for fake offers.

- **Increased Costs and Resource Usage:** Online platforms must spend a lot of money and resources to fight spam. Filtering spam requires powerful servers, which can slow down websites. Companies also need to invest in cybersecurity tools and hire security teams. Sometimes, human moderators are needed to check content, adding to operational costs.

- **Loss of Trust in Online Platforms:** When users see too much spam, they start losing trust in online platforms. For example, an e-commerce site with too many fake reviews may lose customers. On social media, users may stop engaging with posts if they frequently see spam. Platforms that fail to control spam may lose credibility, causing users to switch to more secure alternatives.

4. Challenges in Detecting Spam

Detecting spam is a difficult task because spammers constantly find new ways to avoid detection. Automated systems must be smart enough to recognize spam while ensuring that genuine messages are not mistakenly blocked. Some of the major challenges in spam detection include:

- **Evolving Tactics:** Spammers constantly adapt their tactics to evade detection systems. They often alter the wording of their messages, attach different types of files, or set up fake accounts to distribute spam. Some even use special characters, extra spaces, or symbols to deceive spam filters into classifying their messages as legitimate. Therefore, spam detection techniques need to be regularly updated to counter these evolving strategies.

- **False Positives:** One of the biggest challenges in spam detection is false positives, which happen when a genuine message is wrongly identified as spam. For example, an important business email or a message from a new contact may be mistakenly classified as spam, causing communication issues. If a spam filter is too strict, it might block useful messages, leading to frustration for users. On the other hand, if it is too lenient, spam messages may flood the platform.

- **Volume of Content:** Online platforms receive an enormous amount of content every day, including emails, social media posts, and comments. Manually checking all of this content for spam is impossible. Automated spam detection systems must be efficient enough to handle large-scale data without slowing down the platform. The challenge is to develop models that can process huge amounts of information quickly while maintaining high accuracy.

- **Language and Context Understanding:** Spam messages often use misleading language, unusual symbols, or vague phrases that can make detection difficult. Some spammers try to make their messages look like regular conversations, making it harder to tell if they are spam. Simply looking for specific words is not enough; a good spam detection system must also understand the meaning and intent behind the text. This requires advanced techniques like natural language processing (NLP) and machine learning to accurately identify spam while reducing errors.

5. Types of Spam in Digital Platforms

Spam appears in different forms online and can cause security risks and inconvenience for users. Below are some common types of spam found on digital platforms:

- **Email Spam:** Email spam refers to unwanted bulk messages, often used for advertisements, scams, or phishing attacks. Phishing emails pretend to be from trusted companies to steal personal details, while scam emails lure users with fake lottery wins, job offers, or investment deals to trick them into sending money. Additionally, promotional spam consists of unsolicited emails advertising products or services without user permission.

- **Social Media Spam:** Social media spam appears on platforms like Facebook, Twitter, and Instagram in different forms. Fake accounts are bots or fraudulent profiles used to spread false information or scams. Comment spam includes unwanted promotional messages under posts, often containing harmful links. Message spam involves unsolicited direct messages with scam offers or dangerous links, tricking users into clicking on them.

- **Search Engine Spam (SEO Spam):** Search engine spam refers to unfair techniques used by websites to rank higher in search results. Keyword stuffing involves overusing keywords unnaturally to manipulate rankings. Link farming creates fake backlinks to make a site appear more credible. Hidden text and cloaking show different content to search engines than what real users see, misleading both search engines and visitors.

- **Web Forum and Blog Spam:** Web forum and blog spam occurs when spammers post irrelevant or harmful content in online discussions. This includes comment spam, where promotional or misleading messages with harmful links are posted under blog articles or forum threads. Fake reviews are used to deceive customers by giving false positive or negative feedback about products or services. Link spam involves posting scam links in forums and blog comments to direct users to fraudulent websites.

- **SMS and Messaging Spam:** Spam is also common in SMS and messaging apps like WhatsApp and Telegram. Smishing (SMS phishing) involves fake text messages designed to steal personal information. Scam messages include fake alerts about lottery wins, account updates, or deliveries to trick users. Promotional spam refers to unwanted messages advertising products or scams without user consent.

- **Video and Streaming Spam:** On platforms like YouTube and streaming services, spam appears in different ways. Spam videos are misleading videos that promote scams. Fake live streams falsely claim to offer giveaways but redirect users to scam websites. Clickbait titles and thumbnails use misleading images and titles to attract viewers but provide irrelevant or deceptive content.

- **Voice and Robocall Spam:** Automated calls, known as robocalls, are often used for scams. Telemarketing spam includes unwanted calls promoting fake products or services. IRS/tax scams involve calls pretending to be from tax authorities, demanding payment. Tech support scams trick users by claiming their devices have a virus and offering fake technical assistance.

- **Cryptocurrency and Investment Spam:** Cryptocurrency scams trick users into losing money or account access. Fake airdrops and giveaways promise free cryptocurrency but steal account details. Ponzi schemes are fraudulent investment programs where old investors are paid using money from new investors. Phishing attacks use fake login pages to steal cryptocurrency wallet credentials.

- **Fake Apps and Software Spam:** Some spammers create harmful apps that steal user data or spread malware. Fake antivirus software pretends to remove viruses but actually installs harmful programs. Adware apps bombard users with excessive unwanted ads. Data-harvesting apps secretly collect and sell personal information without the user's knowledge.

6. Fake Account Detection

Social media and online platforms have made communication easier, but they have also led to the rise of fake accounts. These accounts are often created for harmful purposes like spreading false information, scamming people, or influencing opinions. Detecting fake accounts is important to keep online spaces safe and trustworthy.

Fake accounts come in different types. Bot accounts are automated and perform repetitive actions like liking and sharing posts. Impersonation accounts pretend to be real people to mislead others. Scam accounts trick users by promoting fake offers or phishing links. These accounts can be used to manipulate social media, spread viruses, or steal personal data.

Finding fake accounts is not easy because some of them act like real users. Basic detection methods look for suspicious activities like too many posts in a short time, missing profile details, or repetitive messages. However, advanced fake accounts can avoid these checks by copying real human behavior.

To improve detection, machine learning and artificial intelligence (AI) are used. These technologies analyze user behavior, connections, and content to spot unusual activities, such as a sudden increase in followers or strange interaction patterns. Natural language processing (NLP) helps detect fake messages, spam comments, and misleading reviews.

Another method to detect fake accounts is image and biometric analysis. Many fake profiles use stolen or computer-generated pictures. Advanced systems can check profile pictures through facial recognition and reverse image search to verify if they are real.

Even with advanced detection techniques, cybercriminals keep finding new ways to create fake accounts. To stay ahead, online platforms need to regularly update their detection systems and apply stricter security measures like two-factor authentication and identity verification. Working together with cybersecurity experts and government agencies can help reduce the number of fake accounts.

In short, detecting fake accounts is important for keeping digital platforms safe. By using AI, machine learning, and behavior analysis, online platforms can find and remove fake accounts, making the internet a safer place for everyone.

7. Challenges in Detecting Fake Accounts

Detecting fake accounts is a tough task because cybercriminals constantly find new ways to avoid detection. As security improves, scammers develop smarter techniques to create and use fake profiles. Below are some key challenges in identifying fake accounts:

- **Smart Bots Acting Like Real Users:** Some fake accounts use advanced bots that behave like real people by liking posts, commenting, and sharing content. These bots can trick basic detection systems that look for simple patterns.
- **Stolen or AI-Generated Profile Pictures:** Fake accounts often use stolen photos or AI-generated images, making it difficult to detect them. Reverse image searches can help, but AI-generated pictures are becoming more realistic and harder to recognize.
- **Changing Strategies to Avoid Detection:** Scammers constantly update their methods to stay hidden. They may wait before posting, spread their activity over time, or mix real and fake interactions to look more authentic.
- **Fake Friends and Followers:** Some fake accounts gain followers and interact with real users to appear genuine. This makes it challenging to identify them based on their social connections.
- **Tricky Use of Language and Content:** Fake accounts use well-written messages and AI-generated text that sound like real people. This makes it hard for traditional spam detection tools to spot them.
- **Difficult Identity Verification:** Many platforms require phone or email verification, but scammers use temporary emails and virtual phone numbers to create multiple fake accounts. Stronger methods like biometric verification can help, but they also raise privacy concerns.
- **Huge Number of Accounts to Check:** Social media and online platforms have millions of users, making it impossible to manually check every account. Automated systems must be accurate, but they also risk banning real users by mistake.
- **Privacy and Security Issues:** Detecting fake accounts requires analyzing user data, which raises privacy concerns. Platforms must find a balance between security and protecting user privacy.
- **Fake Accounts on Multiple Platforms:** Scammers often create fake accounts on different websites. If they get banned on one platform, they can easily continue their activities elsewhere.
- **No Universal Detection Method:** Each platform uses different methods to detect fake accounts, and there is no single standard for identifying them. This makes it easier for scammers to find loopholes and continue their activities.

8. Conclusion

This research highlights the importance of advanced machine learning techniques in effectively detecting spam comments and fake accounts on popular online platforms such as YouTube and Facebook. By incorporating behavior analysis, natural language processing (NLP), and user metadata evaluation, the proposed system successfully differentiates between genuine and fraudulent activities with high accuracy. The study also addresses key challenges, including the adaptability of spammers, high content volumes, and evolving cybercriminal strategies. The findings demonstrate that machine learning models, when regularly trained and updated, can significantly reduce spam-related disruptions and improve the security of social platforms. The implementation of a user-friendly application further supports real-time detection, making it practical for both users and administrators. Future work may involve integrating deep learning models and exploring cross-platform detection systems to further strengthen online security against spam and fake accounts.

References

[1] G. K. Soni, A. Rawat, S. Jain and S. K. Sharma, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm with LSB Watermark Technique", Springer Smart Systems and IoT: Innovations in Computing. Smart Innovation, Systems and Technologies, Vol. 141, pp. 483-492, 2020.

[2] S. A. Saiyed, N. Sharma, H. Kaushik, P. Jain, G. K. Soni and R. Joshi, "Transforming portfolio management with AI and ML: shaping investor perceptions and the future of the Indian investment sector," Parul University International Conference on Engineering and Tec

[3] H. Kaushik. "Artificial Intelligence in Healthcare: A Review". International Journal of Engineering Trends and Applications (IJETA), Vol. 11, Issue. 6, pp. 58-61, 2024.

[4] N. K. Tiwari and H. Arora, "Sentiment Analysis and Forecasting for Improved Business Performance in E-Commerce using Machine Learning Algorithms," 2025 International Conference on Electronics and Renewable Systems (ICEARS), pp. 1487-1491, 2025.

[5] V. Joshi, S. Patel, R. Agarwal and H. Arora, "Sentiments Analysis using Machine Learning Algorithms," IEEE 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), pp. 1425-1429, 2023.

[6] S. K. Shakya, Dr. R. Misra, "Face Recognition Attendance System, Smart Learning, College Enquiry Using AI Chat-Bot", International Conference on Recent Trends in Engineering & Technology (ICRTET-2023), pp. 164-170, 2023.

[7] H. Kaushik. "Artificial Intelligence: Recent Advances, Challenges, and Future Directions". International Journal of Engineering Trends and Applications (IJETA) Vol. 12(2), pp. 7-13, 2025.

[8] R. Joshi, M. Farhan, U. Sharma, S. Bhatt, "Unlocking Human Communication: A Journey through Natural Language Processing", International Journal of Engineering Trends and Applications (IJETA), Vol. 11, Issue. 3, pp. 245-250, 2024.

[9] H. Sharma, N. Seth, H. Kaushik, K. Sharma, "A comparitive analysis for Genetic Disease Detection Accuracy Through Machine Learning Models on Datasets", International Journal of Enhanced Research in Management & Computer Applications, Vol. 13, Issue. 8, 2024.

[10] H. Kaushik, H. Arora, R. Joshi, K. Sharma, M. Mehra and P. K. Sharma, "Digital Image Security using Hybrid Model of Steganography and Cryptography," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 1009-1012

[11] R. Misra, "A Novel Approach to Enhanced Digital Image Encryption Using the RSA Algorithm", International Conference on Engineering & Design (ICED), 2021.

[12] H. Kaushik, K. D Gupta, "Code Clone Detection: An Empirical Study of Techniques for Software Engineering Practice", Lampyrid: The Journal of Bioluminescent Beetle Research, Vol. 13, pp. 61-72, 2023.

[13] J. Dabass, K. Kanhaiya, M. Choubisa, K. Gautam, "Background Intelligence for Games: A Survey", Global Journal on Innovation, Opportunities and Challenges in AAI and Machine Learning, Vol. 6, Issue. 1, pp. 11-22, 2022.

[14] H. Arora, G. K. Soni, R. K. Kushwaha and P. Prasoon, "Digital Image Security Based on the Hybrid Model of Image Hiding and Encryption," IEEE 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1153-1157, 2021.

[15] S. Pathak, S. Tiwari, K. Gautam, J. Joshi, "A Review on Democratization of Machine Learning In Cloud", International Journal of Engineering Research and Generic Science, Vol. 4, Issue. 6, pp. 62-67, 2018.

[16] K. Gautam, M. Dubey, N. Jain, "Face Detection and Recognition for Patient", International Journal of Biomedical Engineering, Vol. 8, Issue. 2, pp. 1-7, 2022.

[17] V. Singh, M. Choubisa, G. K. Soni, "Enhanced Image Steganography Technique for Hiding Multiple Images in an Image Using LSB Technique", TEST Engineering Management, vol. 83, pp. 30561-30565, May-June 2020.

[18] P. Jain, R. Joshi, "Bridging the Divide Between Human Language and Machine Comprehension", International Conference on Recent Trends in Engineering & Technology (ICRTET 2023), 2023.

[19] Dr. Himanshu Arora, Gaurav Kumar Soni, Deepti Arora, "Analysis and Performance Overview of RSA Algorithm", International Journal of Emerging Technology and Advanced Engineering, Vol. 8, pp. 9-12, 2018.

[20] K.Kanhaiya, A. K. Sharma, K. Gautam, P. S. Rathore, "AI Enabled-Information Retrival Engine (AI-IRE) in Legal Services: An Expert-Annotated NLP for Legal Judgements", 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 2023.

[21] G. K. Soni, H. Arora, B. Jain, "A Novel Image Encryption Technique Using Arnold Transform and Asymmetric RSA Algorithm", Springer International Conference on Artificial Intelligence: Advances and Applications 2019 Algorithm for Intelligence System, pp. 83-90, 2020.

[22] K. Gautam, S. K. Yadav, K. Kanhaiya, S. Sharma, "Hybrid Software Development Model Outcomes for In-House IT Team in the Manufacturing Industry", International Journal of Information

Technology Insights & Transformations (Eureka Journals), Vol. 6, Issue. 1, pp. 1-10, 2022.

[23] H. Sharma N. Seth, H. Kaushik, K. Sharma, "A comparitive analysis for Genetic Disease Detection Accuracy Through Machine Learning Models on Datasets", International Journal of Enhanced Research in Management & Computer Applications, Vol. 13, Issue. 8, 2024.

[24] R. Joshi, A. Maritammanavar, "Deep Learning Architectures and Applications: A Comprehensive Survey", International Conference on Recent Trends in Engineering & Technology (ICRTET 2023), pp. 1-5, 2023.

[25] H. Kaushik, K. D. Gupta, "Machine learning based framework for semantic clone detection", Recent Advances in Sciences, Engineering, Information Technology & Management, pp. 52-58, 2025.