International Journal of Recent Research and Review, Vol. XVIII, Issue 1, March 2025
ISSN 2277 – 8322

Spam and Fake Account Detection Using Machine Learning: A Review

Vivek Jethani

M.Tech Student, Department of professor CSE,

Arya College of Engineering and IT, Jaipur, Rajasthan, India

vivekkrjethani@gmail.com,

Dr. Vibhakar Pathak

Professor, Department of CS and IT,

Arya College of Engineering and IT, Jaipur, Rajasthan, India.

vibhakar@aryacollege.in

Abstract: The proliferation of spam content and fake accounts on digital platforms presents significant challenges to maintaining secure and trustworthy online environments. This review paper provides a comprehensive overview of recent advancements in spam and fake account detection using machine learning (ML) techniques. It explores various data-driven approaches, algorithms, datasets, and evaluation metrics employed in combating malicious online behavior. By analyzing trends, challenges, and future directions, this paper aims to guide researchers and practitioners in developing more robust and intelligent detection systems.

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

1. Introduction

As online platforms continue to grow, ensuring the authenticity of user interactions has become increasingly important. Social media websites like YouTube and Facebook are often targeted by spammers and fraudulent users who spread misleading information, advertisements, and harmful content. Additionally, fake accounts are created to manipulate engagement, promote scams, or even commit cybercrimes. These activities not only disrupt genuine interactions but also pose security risks to users and platforms. Detecting and preventing such activities is crucial for maintaining a safe and trustworthy online environment.

Traditionally, platforms have used rule-based methods to detect spam and fake accounts. These methods rely on predefined rules, such as identifying specific keywords, blacklisting certain users, or flagging repetitive behavior. However, these techniques have significant limitations because spammers and fraudsters constantly change their strategies to avoid detection. As a result, machine learning algorithms have emerged as a more effective solution for identifying spam and fake accounts.

Machine learning models can analyze large amounts of data and identify hidden patterns that differentiate genuine users from fraudulent ones. Some of the most commonly used machine learning algorithms for this task include: Logistic Regression, Random Forest, Support Vector Machine (SVM), Naïve Bayes and more.

These models learn from past data and improve over time, making them more effective in detecting new types of spam and fake accounts. By examining factors such as text content, user behavior, and account details, machine learning techniques can accurately classify comments as spam or genuine and determine whether a Facebook account is real or fake.

This study focuses on developing a machine learning-based system to detect spam comments on YouTube and identify fake accounts on Facebook. An interactive application has been designed to help users easily check whether a comment is spam or an account is fake. Through experimental analysis, the effectiveness of different machine learning models has been tested, showing high accuracy in detecting fraudulent activities.

2. Spam Detection and Classification

The growth of the internet and online platforms has brought many benefits, but it has also led to a significant increase in unwanted, harmful, or irrelevant content commonly referred to as spam. Spam is a major issue across various online services like email, social media platforms, and websites. It includes messages or content that is sent in bulk, usually for malicious purposes or to promote something unrelated to the user's interests. Understanding and addressing spam is crucial for maintaining a safe, efficient, and enjoyable online experience.

Spam refers to any unwanted or unsolicited content that is sent to a large number of people with the aim of promoting products, services, or spreading misinformation. While spam is often associated with email, it also appears on social media platforms, messaging apps, comment sections, and even through 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 frequently change their strategies to bypass detection systems. They may modify the wording of their messages, use different file attachments, or create fake accounts to spread spam. Some spammers use special characters, spaces, or symbols to trick filters into thinking the message is legitimate. As a result, spam detection methods must constantly be updated to keep up with new tricks used by attackers.

- **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

Machine learning has brought major improvements in detecting spam and fake accounts, offering smarter and more scalable solutions than traditional rule-based methods. As online threats continue to evolve, these intelligent systems help identify complex patterns and suspicious behaviors more effectively.

Techniques such as Support Vector Machines (SVM), Random Forest, Naïve Bayes, and Natural Language Processing have shown great potential across various platforms like email, social media, and forums. However, challenges such as false positives, data privacy concerns, and rapidly changing attack strategies still remain.

To stay ahead of cybercriminals, ongoing research is needed in advanced areas like deep learning, adversarial training, and multi-platform detection. A collaborative approach involving developers, cybersecurity experts, and policymakers will be key to building more secure and reliable online spaces. With continuous innovation, machine learning will play an even greater role in protecting digital platforms from spam and fake accounts.

References

[1]. A. Iqbal, M. Younas, S. Iftikhar, F. Fatima, R. Saleem, "Spam detection using hybrid model on fusion of spammer behavior and linguistics features", Egyptian Informatics Journal, Vol. 29, pp. 1-10, 2025.

[2]. Qazi, N. Hasan, R. Mao, M. Elhag Mohamed Abo, S. Kumar Dey and G. Hardaker, "Machine Learning-Based Opinion Spam Detection: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 143485-143499, 2024.

[3]. S. Rastogi, R. Sambyal, P. Tyagi and R. Kushwaha, "Multinomial Naive Bayes Classification AlgorithmBased Robust Spam Detection System," IEEE OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, pp. 1-5, 2024.

[4]. 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.

[5]. 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.

[6]. 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.

[7]. G. Nasreen, M. M. Khan, M. Younus, B. Zafar, M. K. Hanif, "Email spam detection by deep learning models using novel feature selection technique and BERT", Egyptian Informatics Journal, Vol. 26, pp. 1-11, 2024.

[8]. H. Arora, M. Kumar, T. Rasool and P. Panchal, "Facial and Emotional Identification using Artificial Intelligence," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1025-1030, 2022.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. 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.

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

[14]. 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.

[15]. 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.

[16]. A. S. Xiao, Q. Liang, "Spam detection for Youtube video comments using machine learning approaches", Machine Learning with Applications, Vol. 16, pp. 1-9, 2024.

[17]. N. Bandzovic, A. Pasic, D. Mehanovic and A. Dzelihodzic, "Exploring and Analyzing Spam Messages: A Comprehensive Study Using Python, Natural Language Processing and Machine Learning Models," IEEE 28th International Conference on Information Technology (IT), pp. 1-5, 2024.

[18]. 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.

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

[20]. 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.

[21]. 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.

[22]. S. M. Nagare, P. P. Dapke, S. A. Quadri, S. B. Bandal and M. R. Baheti, "Short Message Service (SMS) Mobile Spam Detection using Naïve Bayes," IEEE 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), pp. 67-70, 2024.