International Journal of Engineering Trends and Applications (IJETA) – Volume 12 Issue 6, Nov-Dec 2025

RESEARCH ARTICLE OPEN ACCESS

A Comprehensive Review of Machine Learning Techniques for Detecting Spam Comments on YouTube

Dinesh Kumar Tak, Amit Kumar Sharma
Department of Computer Science and Engineering, Rajasthan Institute of Engineering and Technology, Jaipur, Rajasthan, India

Abstract:
YouTube has emerged as one of the most influential and widely used video-sharing platforms, allowing billions of users to interact through comments. However, this interactive space has become a target for spammers disseminating malicious links, scams, adult content promotions, and misleading advertisements. Manual moderation is inefficient due to the scale of user activity. Machine Learning (ML) offers automated and intelligent spam detection solutions by analyzing behavioral and textual features of comments. This review paper presents a comprehensive analysis of ML-based techniques, datasets, feature engineering strategies, evaluation metrics, and state-of-the-art frameworks used for detecting spam comments on YouTube. Additionally, the paper highlights challenges, open research gaps, and future directions to support the development of more robust and adaptable spam detection systems.

Keywords: Social Media, Machine Learning, Spam Detection, YouTube, Algorithms, Digital Data, Artificial Intelligence.

1. Introduction

The rapid expansion of digital communication platforms has revolutionized the way users interact, share information, and express opinions globally. While this online ecosystem has created countless opportunities for learning and collaboration, it has also introduced significant challenges in maintaining the authenticity and reliability of user-generated content. Among various social media and content-sharing platforms, YouTube stands out as one of the most widely used and influential networks, attracting millions of daily users. However, its open commenting system has made it vulnerable to the widespread issue of spam comments.

Spam comments on YouTube commonly include misleading information, promotional advertisements, malicious links, phishing attempts, and content generated solely to manipulate engagement metrics. These activities not only disrupt meaningful user interaction but also dilute content quality, misguide viewers, and damage the platform's credibility. In severe cases, spam can even expose users to cyber threats such as financial scams and malware attacks. Therefore, identifying and eliminating spam comments is crucial to ensure a safe, trustworthy, and engaging digital environment.

Historically, YouTube and similar platforms relied on rule-based spam detection techniques, which filtered comments using predefined keyword lists, repetitive posting patterns, or blacklisted accounts. While these methods provided a basic level of protection, they lack adaptability and struggle to keep pace with rapidly evolving spam strategies. Spammers frequently alter language patterns, use disguised links, and exploit loopholes to avoid detection. As a result, rule-based approaches are neither scalable nor effective for modern spam detection requirements.

To address these limitations, machine learning has emerged as a robust and intelligent solution. Unlike static rule-based systems, machine learning algorithms can learn from large datasets, analyze hidden patterns, and automatically classify comments as spam or non-spam with high accuracy. Models such as

Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naïve Bayes have proven particularly powerful for text classification tasks due to their ability to extract meaningful features from textual data.

A machine learning-based spam detection system continuously adapts to new trends by training on updated datasets, which allows it to recognize emerging spam behaviors that traditional filters cannot detect. By leveraging features such as linguistic structure, keyword frequency, user behavior, posting patterns, and contextual relevance, these models provide a dynamic, scalable, and highly accurate approach to maintaining the authenticity of YouTube comment sections.

The integration of machine learning into spam detection represents a significant advancement toward securing digital communication. A well-designed machine learning model can not only reduce the volume of harmful and irrelevant comments but also contribute to a safer and more enjoyable environment for users, content creators, and platform administrators.

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.

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, YuTube 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. Conclusion

Machine Learning has proven to be highly effective in detecting spam comments on YouTube, exceeding traditional methods in scalability, adaptability, and accuracy. Deep Learning models such as CNNs, LSTMs, and transformers now dominate research due to their semantic understanding of language. However, evolving spam patterns, limited labeled datasets, multilingual diversity, and adversarial threats pose persistent challenges. Future research focusing on hybrid models, adversarial defense, explainable AI, and real-time frameworks will enhance YouTube's content moderation ecosystem and ensure safer user engagement.

REFERENCES

[1]. V. K. Jethani, Dr. V. Pathak, Dr. V. Shrivastava, "A Machine Learning-Based Approach for Spam Detection and Fake Account Identification on Social Media Platforms", International Journal of Engineering Trends and Applications (IJETA)- Vol. 12 Issue. 4, Jul-Aug 2025, pp. 90-98,2025

[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]. 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.

[5]. 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.

[6]. A. Maheshwari, R. Ajmera and D. K. Dharamdasani, "Unmasking Embedded Text: A Deep Dive into Scene Image Analysis," 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), pp. 1403-1408, 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]. 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.

[9]. 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

[10]. H. Sharma and R. Ajmera, "Comprehensive review and analysis of elderly fall detection system using machine learning," Tuijin Jishu/Journal of Propulsion Technology, vol. 44, no. 5, 2023.

[11]. R. Ajmera and N. Saxena, "Face detection in digital images using color spaces and edge detection techniques," Int. J. of Advanced Research in

Computer Science and Software Engineering, vol. 3, no. 6, pp. 718–725, Jun. 2013.
[12]. R. Misra, "A Novel Approach to Enhanced Digital Image Encryption Using the RSA Algorithm", International Conference on Engineering & Design (ICED), 2021.
[13]. 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.
[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]. G. Sharma, N. Hemrajani, S. Sharma, A. Upadhyay, Y. Bhardwaj, and A. Kumar, "Data management framework for IoT edge-cloud architecture for resource-constrained IoT application," Journal of Discrete Mathematical Sciences and Cryptography, vol. 25, no. 4, pp. 1093–1103, 2022.
[16]. 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.
[17]. 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.
[18]. V. Jethani, Dr. V. Pathak, "Spam and Fake Account Detection Using Machine Learning: A Review", International Journal of Recent Research and Review (IJRRR), Vol. XVIII, Issue. 1, pp. 230-240, 2025.
[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]. H. Sharma and R. Ajmera, "Comprehensive review and analysis on machine learning based Twitter opinion mining framework," Tuijin Jishu/Journal of Propulsion Technology, vol. 44, no. 5, 2023.
[22]. S. Thapar, G. K. Soni, H. Kaushik, R. Singh, S. Bisht and S. K. Bansal, "A Comparative Machine Learning Framework for Detecting Fake Accounts on Facebook," 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 1567-1571, 2025.
[23]. M. Kumar, R. Ajmera, and D. Kumar, "Statistical analysis and accuracy assessment of improved machine learning based opinion mining framework," Advances in Nonlinear Variational Inequalities, vol. 27, no. 1, 2024.
[24]. 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.
[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.
[26]. V. K. Jethani, Dr. V. Pathak, Dr. V. Shrivastava, "A Scalable Multi Modal Machine Learning Framework for

Detecting Spam Content on YouTube and Fake Account on Facebook", IEEE 8th International Conference on Computing Methodologies and Communication (ICCMC 2025), 2025.