RESEARCH ARTICLE OPEN ACCESS

Machine Learning-Based Approach 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:
With the increasing use of social media and content-sharing platforms, the rise of spam comments has become a serious threat to user experience, platform integrity, and information reliability. YouTube, in particular, encounters a substantial volume of malicious and promotional messages intended to mislead users and manipulate engagement. Conventional rule-based and manual detection methods are insufficient against the rapidly evolving nature of spam, highlighting the need for automated and intelligent solutions. This thesis presents a machine learning-based framework for detecting spam comments on YouTube. A user-friendly application was developed to automate the detection process. Four classification algorithms like Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naïve Bayes, were trained and evaluated on a labeled dataset of YouTube comments. Experimental analysis demonstrates that the proposed approach is highly effective, achieving up to 96% accuracy in distinguishing spam from legitimate comments. The findings confirm that machine learning offers a robust and scalable solution for detecting spam content on digital platforms. Future extensions of this work may include real-time spam filtering, integration of deep learning models for enhanced feature extraction, multilingual dataset expansion, and deployment across broader social media ecosystems.

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

1. Introduction

YouTube has become one of the most active social media platforms, enabling global interaction through user comments [1]. However, the openness of this communication space has made it a target for spam including misleading promotions, phishing links, scams, and automated bot-generated content. Such spam not only disrupts genuine conversations but also threatens user safety and undermines platform credibility [3], [11].

Traditional rule-based spam filters are no longer adequate, as spammers constantly modify their tactics to evade detection. Machine Learning (ML) offers a more adaptive and intelligent solution by learning patterns from real-world data and accurately classifying spam comments using linguistic and behavioral features. Algorithms such as Naïve Bayes, Support Vector Machines, Logistic Regression, and Random Forest have shown strong performance in YouTube spam detection tasks.

Therefore, the integration of ML techniques is essential for ensuring secure, trustworthy, and engaging user interactions on YouTube. This review focuses on existing ML approaches, their effectiveness, challenges, and potential future advancements in spam comment detection.

2. Proposed Methodology

The flowchart of the proposed methodology is shown in Figure 1. The following steps are involved in detecting spam on YouTube.

- **Data Collection:** The first step is to gather the required data. In this process, we collect datasets in CSV

format from YouTube. The YouTube dataset is used for spam detection.

- **User Interface Interaction:** In this proposed system, users interact with an application where they can input a comment in a text box. When they click the "Check for Spam" button, the system will analyze the comment and determine whether it is spam or not.
- **Data Preprocessing:** Data preprocessing, also known as data cleaning, is an important step in spam detection. This process involves refining raw text by removing unnecessary information, making it easier for the system to analyze the data accurately.

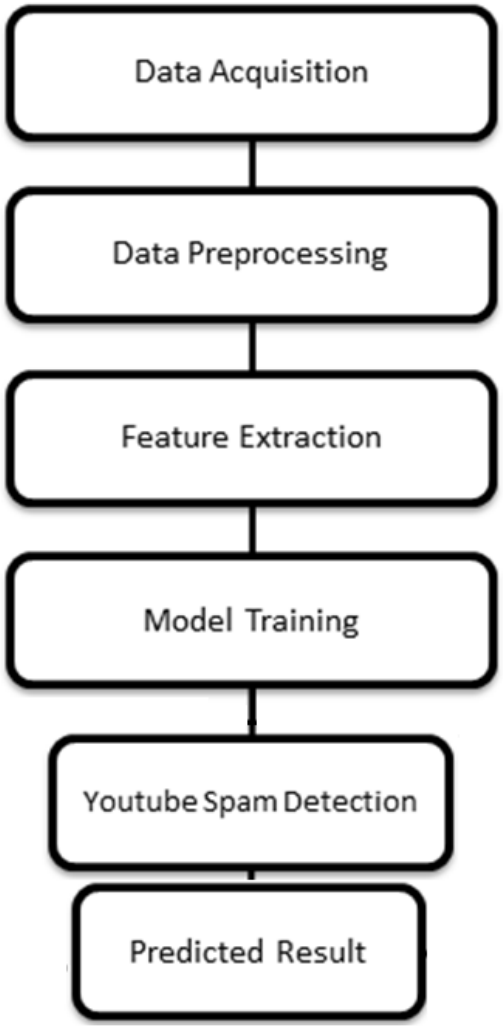

Figure 1: Proposed Methodology for Spam Detection on YouTube Data

- **Feature Extraction:** In this step, the text data is converted into numerical values that machine learning models can understand. The TF-IDF Vectorization method is used to assign importance to words in the dataset. Additionally, N-grams (unigrams and bigrams) are extracted to capture word relationships and improve the accuracy of spam detection.
- **Model Training:** To build an effective detection system, different machine learning models are used. For spam detection on YouTube, we train models using methods such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naïve Bayes. These models help analyze patterns and classify content accurately.

By following these steps, the system aims to improve the detection of spam on YouTube and identify fake accounts on Facebook effectively.

3. Results and Discussion

Based on the outcomes of this study, created an interactive application to enhance usability and simplifying process, this application provides a rectangular input box which enable users to input their comment for check spam detection and under the rectangular input box there are one rectangular check for spam button as show in Figure 2.

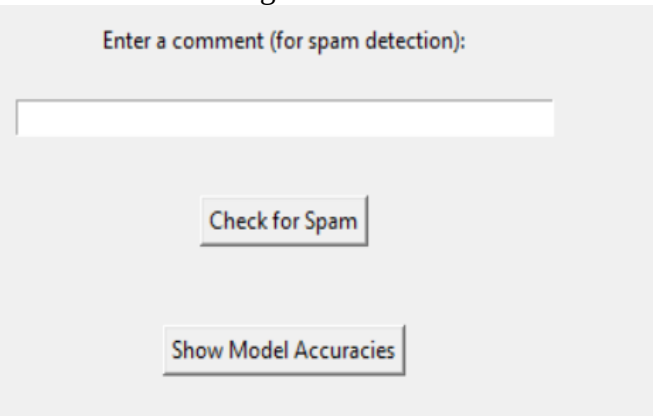

Figure 2: Spam detection application interface

For spam detection, first user need insert a comment that will be analyzed further that is

spam or not. In Figure 3 shown the entered the comment "I sell this laptop" in the input section, the background spam detection application processes the text using a machine learning model. The model analyzes various features to determine whether the comment is spam or not.

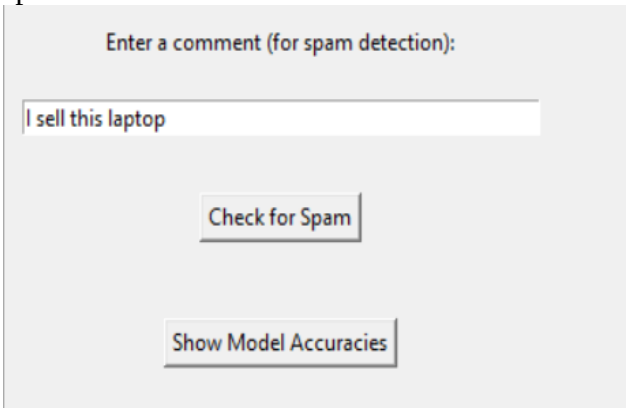

Figure 3: Enter a comment ("I sell this laptop") for spam detection

After entering the comment "I sell this laptop", click the Check Spam option as shown in Figure 3. The application then analyzes the comment in the background and generates a prediction result to determine whether the comment is spam or not. Figure 4 shows that the given comment is classified as Not Spam.

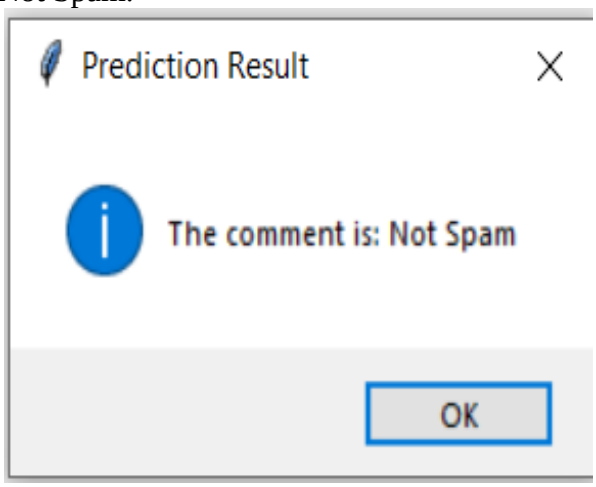

Figure 4: The comment is not spam

The accuracy of the proposed model is illustrated in Figure 5, which presents the performance of various machine learning algorithms applied to the dataset. The model was trained for YouTube spam detection, utilizing multiple classification techniques to evaluate their effectiveness.

For YouTube spam detection, the accuracy varied slightly among the classifiers. Logistic Regression demonstrated the highest accuracy at 96%, followed by Random Forest at 95%. Support Vector Machine (SVM) achieved an accuracy of 92%, while Naïve Bayes performed the lowest at 90%. These variations indicate that some models were more effective than others in identifying spam content on YouTube, potentially due to differences in how each algorithm processes textual and metadata-based features.

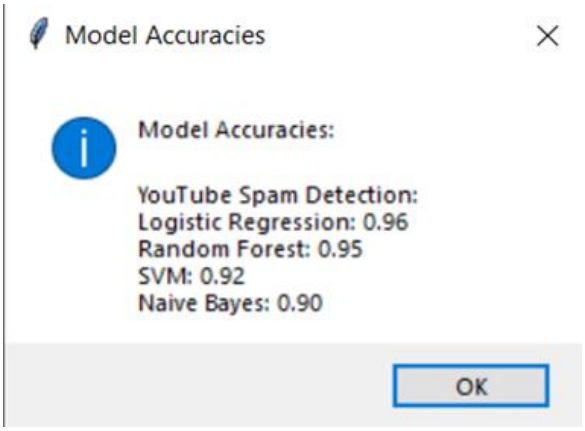

Figure 5: Model accuracies for spam detection on YouTube data

4. Conclusion

This study presents an effective machine learning-based approach for detecting spam comments on YouTube. A user-friendly application was developed to simplify the detection process and enable rapid verification of comments. The experimental analysis demonstrated that machine learning algorithms can accurately distinguish between spam and genuine comments based on textual features.

Among the evaluated models, Logistic Regression achieved the highest accuracy (96%), followed by Random Forest (95%). Support Vector Machine (SVM) attained 92% accuracy, whereas Naïve Bayes recorded 90%. These results indicate that Logistic Regression and Random Forest performed more efficiently in identifying YouTube spam, likely due to their ability to capture relevant linguistic and statistical patterns in user comments.

Overall, the research confirms that machine learning is highly effective for YouTube spam detection and can significantly improve the reliability and safety of user interactions on the platform.

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]. 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.

[4]. 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.

[5]. 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.

[6]. 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.

[7]. 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.

[8]. 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.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. 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.