International Journal of Intelligent Computing Systems

February-2026 | IJICS | Volume-2 Issue-1 | ISSN: 3107-7218

MACHINE LEARNING TECHNIQUES FOR AUTOMATED SPAM DETECTION IN YOUTUBE COMMENTS

NAGARJUNA VANKAYALA, M. Tech is a student of LINGAYAS INSTITUTE OF MANAGEMENT AND TECHNOLOGY, Madalavarigudem-521212, Andhra Pradesh, India

DR. K N VENKATA RATNA KUMAR, Professor Department of Computer Science & Engineering, LINGAYAS INSTITUTE OF MANAGEMENT AND TECHNOLOGY, Madalavarigudem-521212 Andhra Pradesh, India.

ABSTRACT

The exponential growth of user-generated video platforms such as YouTube has transformed the way people communicate, share opinions, and consume digital content. Among the various interaction mechanisms provided by YouTube, the comment section plays a crucial role in facilitating engagement between creators and viewers. However, the openness of this feature has also led to a significant rise in spam comments, including promotional advertisements, phishing links, misleading information, and bot-generated messages. Manual moderation of such content is impractical due to the massive volume of comments posted every minute. This paper proposes a comprehensive supervised machine learning-based framework for automatically detecting and filtering spam comments in YouTube comment sections. A publicly available dataset containing labeled YouTube comments is used for experimentation. Multiple classification algorithms, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Artificial Neural Network (ANN), are implemented and evaluated. Extensive preprocessing and feature engineering techniques are applied to extract meaningful patterns from textual data. Experimental results demonstrate that SVM outperforms other models in terms of accuracy, precision, and F1-score. The proposed system effectively reduces the burden of human moderators while ensuring reliable and scalable spam detection.

Keywords—YouTube comments, Spam filtering, Supervised machine learning, Text classification, Natural Language Processing, Support Vector Machine.

I. INTRODUCTION

Online video-sharing platforms have become an integral part of modern digital communication. YouTube, being one of the largest video-sharing platforms globally, hosts millions of videos and attracts billions of users daily. Users interact with video content not only by viewing and liking but also through comments, which serve as a medium for feedback, discussion, and opinion sharing. While comment sections enhance user engagement, they are also highly susceptible to spam activities.

Spam comments typically include unsolicited advertisements, deceptive

URLs, repetitive promotional messages, and automated bot content aimed at increasing traffic to external websites or manipulating public perception. Such comments degrade the quality of

discussions, mislead users, and negatively affect the overall user experience. Although YouTube provides basic moderation tools such as keyword filtering and spam reporting, these mechanisms are insufficient to address sophisticated and evolving spam strategies.

Given the enormous scale of data generated on YouTube, automated spam detection systems based on machine learning have become essential. Machine learning techniques can learn patterns from historical data and adapt to new spam behaviors. This paper focuses on the design and evaluation of supervised learning models for classifying YouTube comments as spam or non-spam (ham). The objective is to develop an accurate, efficient, and scalable solution that can support real-time content moderation.

II. LITERATURE SURVEY

Spam detection in online platforms has been widely studied, with researchers proposing various machine learning and data mining approaches.

Several studies have explored the use of traditional classifiers such as Naïve Bayes, Logistic Regression, and Support Vector Machines for text-based spam detection. These approaches rely on statistical properties of textual features and have shown promising results in early spam filtering systems.

Recent research has focused on YouTube-specific spam detection, where algorithms such as SVM, K-Nearest Neighbor (KNN), and Random Forest have been applied. Comparative analyses reveal that SVM generally achieves higher accuracy due to its ability to handle high-dimensional feature spaces effectively.

Other works have extended spam detection to identify unsafe or inappropriate content using deep learning techniques, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Although deep learning models provide improved contextual understanding, they require large datasets and high computational resources.

Overall, existing literature emphasizes the importance of preprocessing, feature selection, and algorithm optimization in achieving reliable spam detection. However, there is still a need for a balanced approach that combines high accuracy with computational efficiency.

III. PROBLEM STATEMENT

The rapid increase in spam comments on YouTube poses serious challenges to content moderation and user trust. Manual moderation is time-consuming, expensive, and ineffective at scale. Existing automated systems struggle to adapt to evolving spam patterns and often generate false positives or false negatives. Therefore, there is a need for an intelligent and adaptive spam detection system that can accurately classify comments while operating efficiently in real-time environments.

IV. EXISTING SYSTEM

Traditional spam detection systems primarily relied on rule-based filtering and Artificial Neural Networks. Rule-based systems use predefined keywords and patterns, which are easy to bypass by spammers. ANN-based systems learn complex relationships but require extensive training time and computational resources.

Disadvantages of the Existing System:

- High computational complexity
- Difficulty in tuning model parameters
- Risk of overfitting and underfitting
- Limited scalability for real-time deployment

V. PROPOSED SYSTEM

This paper proposes a supervised learning-based spam detection framework that evaluates multiple classification algorithms to identify the most effective approach.

Algorithms Used:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- Artificial Neural Network (ANN)

The proposed system emphasizes efficient preprocessing, robust feature extraction, and comparative performance evaluation. By selecting the best-performing classifier, the system aims to achieve high accuracy with reduced computational overhead.

Advantages of the Proposed System:

- Improved detection accuracy
- Faster training and prediction time
- Reduced false positives and false negatives
- Scalable for real-time moderation systems

VI. SYSTEM ARCHITECTURE

The overall system architecture consists of the following components:

1. Data Input Layer – Collects raw YouTube comments.
2. Preprocessing Module – Cleans and normalizes text data.
3. Feature Extraction Module – Converts text into numerical features.
4. Classification Module – Applies machine learning algorithms.
5. Output Layer – Labels comments as spam or ham.

VII. METHODOLOGY

1. Dataset Description

The dataset used in this study is obtained from the UCI Machine Learning Repository. It contains labeled YouTube comments categorized as spam and non-spam. The dataset includes textual content and metadata associated with each comment.

2. Data Preprocessing

Preprocessing is performed to improve data quality and model performance. The steps include:

- Removal of special characters, emojis, and HTML tags
- Conversion of text to lowercase

• Tokenization of text into individual words
• Removal of stopwords
• Handling missing and noisy data

3. Feature Engineering

Several features are extracted from the preprocessed text:

• Keyword-based features (e.g., promotional terms)
• Structural features such as comment length
• Presence of URLs and hyperlinks
• Frequency of uppercase words and punctuation

4. Model Training

The dataset is divided into training and testing subsets using an 80:20 ratio. Each classifier is trained on the training set and evaluated on the testing set. Hyperparameters are tuned using cross-validation.

5. Evaluation Metrics

The performance of the models is evaluated using:

• Accuracy
• Precision
• Recall
• F1-Score
• ROC-AUC

VIII. EXPERIMENTAL RESULTS AND ANALYSIS

The performance of the proposed spam detection system was evaluated using multiple supervised machine learning algorithms, namely Logistic Regression, Decision Tree, Artificial Neural Network (ANN), and Support Vector Machine (SVM). The evaluation was carried out on a labeled YouTube comment dataset, which was divided into training and testing sets in an 80:20 ratio. Standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were used to analyze and compare the effectiveness of each model.

Logistic Regression served as a baseline classifier due to its simplicity and efficiency in binary classification problems. The results show that Logistic Regression achieved consistent and stable performance, demonstrating its ability to capture linear relationships between extracted textual features and comment labels. However, its performance slightly declined when handling complex and non-linear spam patterns.

The Decision Tree classifier provided better interpretability by generating rule-based decisions that explain how comments are classified as spam or non-spam. While Decision Trees achieved competitive accuracy, they exhibited a tendency to overfit the training data, especially when deeper trees were used. This resulted in reduced generalization performance on unseen test data.

The Artificial Neural Network (ANN) demonstrated strong recall performance, indicating its effectiveness in correctly identifying a higher proportion of spam comments. This makes ANN suitable for scenarios where minimizing false negatives is critical. However, ANN required higher computational resources, longer training time, and careful parameter

tuning, which may limit its applicability in real-time moderation systems.

Among all evaluated models, Support Vector Machine (SVM) achieved the best overall performance across most evaluation metrics. SVM effectively handled high-dimensional feature spaces and maintained a strong balance between precision and recall. Its ability to construct an optimal decision boundary enabled it to generalize well on unseen data, making it the most robust and reliable classifier for YouTube spam detection in this study.

The experimental analysis confirms that feature engineering combined with an appropriate classifier significantly improves spam detection accuracy. The results clearly indicate that SVM is the most suitable algorithm for large-scale and real-time spam filtering applications.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	88%	85%	87%	86%
Decision Tree	89%	88%	88%	88%
ANN	91%	90%	92%	91%
SVM	**92%**	**93%**	91%	**92%**

X. CONCLUSION

This paper presented a comprehensive supervised machine learning framework for detecting spam comments in YouTube comment sections. The rapid growth of online user-generated content has made manual moderation ineffective, thereby necessitating intelligent automated solutions. By leveraging machine learning techniques, the proposed system efficiently classifies comments into spam and non-spam categories.

Extensive experimentation was conducted using multiple classification algorithms, including Logistic Regression, Decision Tree, Artificial Neural Network, and Support Vector Machine. The results demonstrate that while all models are capable of detecting spam to a certain extent, Support Vector Machine consistently outperforms other classifiers in terms of accuracy, precision, and F1-score. This highlights the effectiveness of SVM in handling textual data and complex decision boundaries.

The study also emphasizes the importance of data preprocessing and feature extraction in improving classification performance. Proper cleaning, normalization, and selection of relevant features significantly enhance the ability of machine learning models to detect spam patterns accurately.

Overall, the proposed system provides a scalable, efficient, and reliable solution for automated spam filtering. It reduces the burden on human moderators, improves user experience, and contributes to maintaining a trustworthy and meaningful

online discussion environment on large-scale platforms such as YouTube.

XI. FUTURE SCOPE

Although the proposed system achieves promising results, there are several directions in which this work can be further extended and enhanced.

Future research can explore the integration of advanced deep learning models such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) to capture contextual and semantic relationships within comments more effectively. These models can improve detection accuracy, especially for complex and disguised spam content.

Incorporating user behavior analysis and network-based features such as user activity patterns, comment frequency, and interaction graphs can further strengthen spam detection capabilities. Such features can help identify coordinated spam attacks and bot-generated content.

The system can also be extended to detect other forms of harmful content, including hate speech, abusive language, cyberbullying, and misinformation. This would transform the system into a comprehensive content moderation framework rather than a spam-only solution.

Additionally, the proposed model can be deployed as a browser extension, RESTful API, or integrated directly with social media platforms for real-time moderation. Continuous learning mechanisms can be implemented to allow the model to adapt to evolving spam strategies and language patterns.

REFERENCES

[1] A. Aziz, C. F. Mohd Foozy, P. Shamala, and Z. Suradi, "YouTube spam comment detection using support vector machine and k-nearest neighbor," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 2, pp. 612–619, 2018.

[2] N. A. Samsudin, M. F. A. Rasid, and A. R. Ahmad, "Spam detection framework for YouTube comments using Naïve Bayes and logistic regression," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1508–1517, Jun. 2019.

[3] A. Ali and M. Z. Amin, "Spam detection for YouTube comments using machine learning techniques," *International Journal of Current Science*, vol. 12, no. 4, pp. 395–403, Oct. 2022.

[4] G. Airlangga, "Spam detection on YouTube comments using advanced machine learning models: A comparative study," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 2, pp. 500–508, Nov. 2024.

[5] M. S. Sam'an and K. Imaddudin, "Hybrid deep learning model for YouTube spam comment detection," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 3, pp. 3313–3319, Jun. 2024.

[6] H. S. Dutta, M. Chakraborty, and P. Goyal, "Detecting and analyzing collusive entities on YouTube," *arXiv preprint arXiv:2005.05530*, May 2020.

[7] A. Sureka, "Mining user comment activity for detecting forum spammers in YouTube," *arXiv preprint arXiv:1103.5044*, Mar. 2011.

[8] H. Sankar, S. Priyadarshini, and R. Balakrishnan, "Feature selection for comment spam filtering on YouTube," *DergiPark Journal of Engineering Sciences*, 2018.

[9] J. Ramos, "Using TF-IDF to determine word relevance in document queries," *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[10] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *European Conference on Machine Learning*, Springer, Berlin, Heidelberg, pp. 137–142, 1998.

[11] Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, 2016.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA, 2016.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[15] UCI Machine Learning Repository, "YouTube Spam Collection Dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection