International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 02 | Feb - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Machine Learning Base Spam Comments Detection on YouTube

1. Swara Sakhare, 2.Mayuri Salve, 3.Snehal Shinde ,4.Shital Kapadi, 5.MrsP.T Gadakh

Student, Department of Information Technology

6. Mr M.P Bhanddkar

HOD, Department of Information Technology
Matoshree Aasarabai Polytechnic, Eklahare, Nashik, Maharashtra-422105

---***---

Abstract - The project proposes a machine learning-based approach for detecting spam comments on YouTube using Natural Language Processing (NLP) algorithms. As user-generated content continues to grow exponentially, spam comments pose a significant threat to user experience and platform integrity. Leveraging a labeled dataset of comments, the system employs various NLP preprocessing techniques, including tokenization, stop word removal, and stemming, to clean and prepare the text data. Advanced feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings are used to capture semantic nuances in the comments. Machine learning classifiers like Logistic Regression (LR), Support Vector Machines (SVM), and Random Forest (RF) are then applied to classify comments as spam or legitimate. Accuracy, precision, recall, and F1-score are employed to determine the models in order to ensure a good balance between false positives and negatives. This research enhances content moderation capabilities by filtering spam and providing insights into spam trends, contributing to improved online content management and a better user experience on social media platforms.

Index Terms- Machine Learning, Spam Detection, YouTube, Natural Language Processing, NLP, Classification, Text Processing, Feature Extraction.

1. INTRODUCTION

The rapid growth of user-generated content on platforms like YouTube has revolutionized information sharing, but it also brings challenges in maintaining a positive user experience due to the prevalence of spam comments. These comments, often containing irrelevant content, self-promotion, or malicious links, disrupt engagement and platform integrity, making robust detection and filtering methods essential. This research proposes a machine learning-based system that leverages Natural Language Processing (NLP) algorithms to effectively classify spam and legitimate comments. NLP, a subfield of artificial intelligence, enables machines to analyze linguistic features, such as word choice, syntax, and sentiment, providing a foundation for machine learning models to learn from labeled data and predict new instances accurately. By employing a labeled dataset of YouTube comments, our approach includes text preprocessing techniques like tokenization, stemming, and stop-word removal to enhance data quality. Semantic relationships within the text have been recorded using feature extraction techniques that include word embeddings and TF-IDF. We evaluate multiple machine learning algorithms, assessing their performance using accuracy, precision, recall, and F1-score to ensure reliable spam detection. By automating the identification of spam, this system not only improves user experience but also reduces the workload of human moderators, contributing to better content management and discourse quality on social media. This research underscores the value of integrating ML and NLP for spam detection, advancing future solutions for online community moderation.

2. Problem Statement

The rapid expansion of YouTube's user-generated content has led to an increase in spam comments, which decreased user engagement and degraded the caliber of discussions. Accurately distinguishing spam from legitimate interactions poses a significant challenge due to the diversity of language and context, as well as the constantly evolving tactics used by spammers. Traditional moderation approaches are inadequate given the massive data volume and dynamic nature of spam. This necessitates an automated, machine learning-based solution powered by Natural Language Processing (NLP) algorithms, capable of detecting and classifying spam comments in real time. Such a system enhances user experience by fostering more meaningful and constructive interactions on the platform.

3. Literature Review

1) Ahmed et al.'s paper from 2022 offers a thorough examination of the many ML approaches utilized for spam detection in email and IoT platforms, emphasizing how well these approaches work to improve security and communication. The authors systematically review existing algorithms, including supervised and unsupervised learning approaches, and discuss their strengths and limitations in different contexts. Additionally, the paper identifies key research challenges in the field, such as the need for robust datasets, feature extraction techniques, and the adaptation of models to evolving spam tactics. By addressing these challenges, the authors underscore the importance of advancing spam detection systems to ensure better protection against unwanted communications, paving the way for future research in this critical area of cybersecurity.[1]
2) Bacanin et al. (2022) examine the integration of Natural Language Processing (NLP) and machine learning techniques enhanced by swarm intelligence for effective spam email filtering. The study emphasizes the importance of feature extraction from email content and presents a hybrid approach that combines traditional machine learning algorithms with swarm intelligence mechanisms to optimize model performance. The authors demonstrate that this innovative framework significantly improves the accuracy and efficiency of spam detection compared to conventional methods. Furthermore, the paper discusses various experiments and results, showcasing the potential of this approach to adapt to the evolving nature of spam emails. By highlighting the synergy between NLP, machine learning, and swarm intelligence, the study contributes valuable insights into developing more resilient and adaptive spam filtering systems.[2]
3) In their study, Danilchenko, Segal, and Vilenchik (2022) present a novel approach to opinion spam detection by leveraging machine learning and network-based algorithms. The authors focus on the challenges posed by deceptive reviews and the difficulty of distinguishing between genuine and spam opinions on online platforms. By integrating network analysis techniques with traditional machine learning models, the research introduces a framework that captures the relationships between users and their interactions, enhancing the detection of coordinated spam activities. The results demonstrate significant improvements in accuracy and reliability over existing methods, indicating that the proposed approach can effectively identify opinion spam in various contexts. This work not only contributes to the field of spam detection but also offers insights into how network dynamics can be utilized to better understand and combat deceptive practices in online environments.[3]
4) In their research, Ghanem and Erbay (2022) investigate spam detection on social networks by employing deep contextualized word representations, a technique that enhances the understanding of the semantic meaning of words within their context. The study emphasizes the limitations of traditional spam detection methods and highlights the need for advanced techniques that can effectively capture the nuances of language used in social media interactions. By utilizing deep learning models, the authors demonstrate that their approach significantly improves the accuracy of spam classification compared to conventional keyword-based methods. The findings indicate that contextualized word embeddings not only facilitate better feature extraction but also allow for more sophisticated analysis of user-generated content, thus enhancing the overall effectiveness of spam detection systems in social networks. This research contributes to the ongoing efforts to improve the integrity and quality of online interactions by providing a robust framework for identifying spam in increasingly complex social media environments.[4]
5) In their study, Novo-Lourés et al. (2022) address the challenges of spam filtering across multiple channels by enhancing the representation of data to improve classification accuracy. The authors propose a comprehensive framework that incorporates various data representation techniques, enabling the system to effectively capture the distinct characteristics of spam content across different communication platforms, that include emails, social media, and messaging apps. By employing advanced feature extraction methods and integrating multiple

sources of information, the research demonstrates a significant improvement in spam detection systems performance. The findings demonstrate the value of a multi-channel strategy for spam filtering since it enables a more comprehensive comprehension of user behavior and content properties. This paper highlights the need for flexible systems that can handle the complexity of modern digital communication contexts and offers insightful information for the creation of more efficient spam filtering solutions.[5]

6) In the corrections article by Oh (2022), the author revisits the previously proposed YouTube spam comments detection scheme that utilizes a cascaded ensemble machine learning model. This paper addresses specific inaccuracies and clarifies methodological details that enhance the understanding of the original research. The corrections emphasize the importance of precise algorithm implementation and the impact of various model parameters on detection accuracy. By refining the details of the ensemble approach, the author highlights how adjustments can lead to improved performance in identifying spam comments effectively. This work underscores the ongoing need for meticulous validation and transparency in machine learning research, particularly in the context of spam detection on platforms like YouTube, where user engagement and content integrity are paramount. Overall, the article serves to strengthen the foundation of the original study, ensuring that researchers and practitioners can build upon a solid and accurate framework for spam detection in online environments.[6]

7) In their study, Sadid, Young, and Rusli (2022) investigate the effectiveness of spam filtering on user feedback through text classification methods, specifically employing the Multinomial Naïve Bayes algorithm combined with the Term Frequency-Inverse Document Frequency (TF-IDF) technique. The authors focus on the challenges associated with filtering spam in user-generated content, emphasizing the need for accurate classification methods to enhance the quality of feedback in online platforms. Through experiments, the research demonstrates that the combination of Multinomial Naïve Bayes and TF-IDF significantly improves the accuracy of spam detection compared to simpler models. The findings indicate that this approach is particularly effective in capturing the distinctive patterns of spam content, making it a viable solution for managing user feedback across various digital platforms. This work contributes to the broader field of spam detection by providing insights into effective text classification techniques that can be implemented to improve content moderation and enhance user engagement in online communities.[7]

8) In their research, Thapa, Lamichhane, Ma, and Jiao (2021) introduce SpamHD, a novel approach to text spam detection that leverages memory-efficient techniques inspired by hyperdimensional computing. The study addresses the growing need for efficient spam detection methods that can operate on limited computational resources while maintaining high accuracy. By utilizing brain-inspired hyperdimensional computing, the authors propose a framework that encodes text data into high-dimensional representations, allowing for effective classification of spam content without the excessive memory requirements typically associated with traditional machine learning approaches. The experimental results demonstrate that SpamHD not only outperforms conventional spam detection methods in terms of accuracy but also significantly reduces memory usage, making it particularly suitable for deployment in resource-constrained environments. This work contributes to the advancement of spam detection technologies by showcasing the potential of hyperdimensional computing as a powerful tool for enhancing efficiency and performance in text classification tasks.[9]

9) In their paper, Douzi, AlShahwan, Lemoudden, and El Ouahidi (2020) present a hybrid model for email spam detection that integrates various artificial intelligence techniques to enhance classification accuracy. The authors explore the limitations of traditional spam detection methods and propose a comprehensive framework that combines different algorithms, including machine learning classifiers and rule-based approaches. By leveraging this hybrid methodology, the study demonstrates improved performance in accurately identifying spam emails compared to standalone models. The experiments conducted reveal the model's robustness in adapting to diverse email content

and its effectiveness in reducing false positives and negatives. This research contributes to the field of spam detection by highlighting the advantages of a hybrid approach, offering valuable insights for the development of more reliable and efficient spam filtering solutions in email communication systems. In an increasingly digital environment, the results highlight the future potential of artificial intelligence in resolving the enduring problems of spam detection.[9]

10) In their study, Li, Wu, and Wang (2019) investigate comment spam detection by exploring the effectiveness of combining various feature extraction techniques. The authors recognize the complexity of detecting spam comments on online platforms, which often involves nuanced language and varying user behaviors. By systematically analyzing and selecting a diverse set of features, including textual, syntactical, and semantic attributes, the researchers develop a comprehensive model that enhances the accuracy of spam detection. The results of the trials show that the combination of useful features significantly improves spam comment detection while reducing false positive accuracy. This study advances the field of spam detection by focusing illumination on how crucial feature combinations and selection are to developing reliable classification models. The findings highlight the potential for applying similar strategies in other domains of content moderation, emphasizing the need for adaptable and effective approaches in the ongoing battle against spam in digital communication.[10]

4. Proposed System

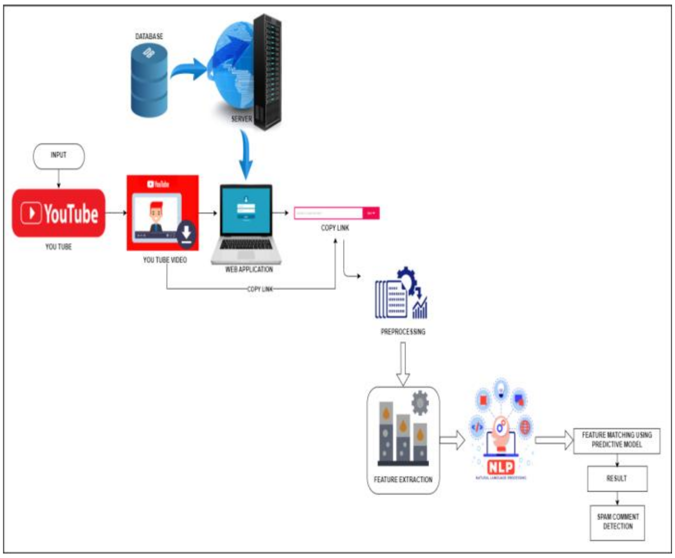

Fig -1: System Architecture

The YouTube machine learning-based spam comment detection system architecture is divided into a number of essential parts, each of which is essential to the system's overall functioning. At the foundation, the architecture begins with the Data Collection Module, which gathers user comments from YouTube using the platform's API. To train and assess the machine learning models, this module collects a varied dataset of comments, containing both spam and valid posts. The Preprocessing Module assumes control after data collection, cleaning, and preparing the text for analysis using NLP techniques including tokenization, stemming, and stop word removal. By accomplishing this, the data will be certain to be in an appropriate format for the feature extraction stage that follows. Following preprocessing, the Feature Extraction Module uses techniques that include TF-IDF or word embeddings to change the textual data into numerical formats so that the comments can be represented in a way that machine learning algorithms can comprehend. To train on the extracted features and differentiate between spam and real comments, use the Model Training and Evaluation Module. After training, the system integrates a Real-Time Detection Module that processes incoming comments, applying the trained models to classify them on the fly. Finally, the User Interface Module presents the results to content creators and moderators, allowing them to manage comments efficiently. The architecture also includes a feedback loop that facilitates continuous learning and model improvement based on user inputs and misclassifications, ensuring the system remains effective against evolving spam tactics.

5. ALGORITHM USED FOR PROPOSED SYSTEM

1) **NLP:** This particular study presents an ML-based system that recognizes YouTube spam comments using NLP techniques. To address the increasing issue of spam in user-generated content, the system uses a labeled dataset of comments and key NLP approaches that involve tokenization, stop word removal, and stemming for effective text preprocessing. Semantic subtleties in comments are captured using feature extraction approaches that include word embeddings and TF-IDF. Various classifiers, that include LR, SVM, and RF, are trained to classify comments as spam or legitimate. Metrics that include accuracy, recall, precision, and F1-score have been employed to determine the model's performance, assuring a balanced detection of false positives and negatives. This system not only improves

spam filtering but also provides insights into spam trends, supporting better content moderation and enhancing user experience on social media platforms like YouTube.

6. Applications

1) Content Creators: YouTube content creators can utilize the system to manage their comment sections effectively, ensuring that discussions remain constructive and relevant, thereby enhancing viewer engagement.
2) Platform Moderators: YouTube's moderation team can leverage the system for real-time monitoring and automatic filtering of spam comments, reducing the manual workload and improving moderation efficiency.
3) Advertising and Brand Safety: Brands and advertisers can use the system to ensure that their advertisements are not associated with spammy or harmful comments, protecting their reputation and ensuring a positive brand image.
4) Community Management Tools: Third-party community management platforms that integrate with YouTube can employ the spam detection system to enhance their moderation features, providing users with better tools for managing online communities.
5) Social Media Analytics: Organizations focused on social media analytics can use the system to analyze user interactions and comment trends, providing insights into audience sentiment and engagement.
6) Research and Development: Academic institutions and researchers can utilize the system for studies related to online behavior, spam detection methodologies, and the impact of spam on user engagement.
7) Customer Support Systems: Businesses that engage with customers via YouTube can use the spam detection system to filter out irrelevant comments, allowing support teams to focus on genuine customer inquiries and feedback.

7. Conclusions

In conclusion, using a machine learning-based system that uses NLP to identify spam comments on YouTube presents an innovative approach to enhancing user interactions on the platform. By automating the identification and removal of spam comments, the system not only reduces the workload for content creators and moderators but also contributes to fostering a healthier online environment that promotes meaningful engagement. The benefits of enhanced accuracy, real-time processing, and system's ability to adapt to evolving spam strategies highlight its importance in addressing contemporary content moderation challenges. As advancements in machine learning and NLP continue, these detection capabilities will only improve, helping platforms like YouTube uphold community standards and encourage constructive conversations among users. Ultimately, this system is a crucial step in ensuring a more positive and enriching experience for both content creators and viewers.

REFERENCES

[1] . Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.

[2] Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine learning techniques for spam detection in email and IoT platforms: Analysis and research challenges. Security and Communication Networks, 2022, 1–19.

[3] Bacanin, N., Zivkovic, M., Stoean, C., Antonijevic, M., Janicijevic, S., Sarac, M., et al. (2022). Application of natural language processing and machine learning J boosted with swarm intelligence for spam email filtering. Mathematics, 10(22), 4173.

[4] Danilchenko, K., Segal, M., & Vilenchik, D. (2022). Opinion spam detection: A new approach using machine learning and network-based algorithms. In Proceedings of the International AAAI Conference on Web and Social Media, vol. 16 (pp. 125–134).

[5] R. Ghanem and H. Erbay, "Spam detection on social networks using deep contextualized word representation", Multimedia Tools and Applications, pp. 1-16, 2022

[6] M. Novo-Lourés, D. Ruano-Ordás, R. Pavón, R. Laza, S. Gómez-Meire and J. R. Méndez, "Enhancing representation in the context of multiple-channel spam filtering", Information Processing & Management, vol. 59, no. 2, pp. 102812, 2022.

[7] H. Oh, "Corrections to "A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model", IEEE Access, vol. 10, pp. 40860-40860, 2022

[8] S. Sadid, J. Young and A. Rusli, "Spam Filtering on User Feedback Via Text Classification Using Multinomial Naïve Bayes and TF-IDF", Ultimatics: Jurnal Teknik Informatika, vol. 13, no. 2, pp. 108-113, 2022.

[9] R. Thapa, B. Lamichhane, D. Ma and X. Jiao, "Spamhd: Memory-efficient text spam detection using brain-inspired hyperdimensional computing", In 2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 84-89, 2021, July.

[10] S. Douzi, F. A. AlShahwan, M. Lemoudden, and B El Ouahidi, "Hybrid Email Spam Detection Model Using Artificial Intelligence", International Journal of Machine Learning and Computing, vol. 10, no. 2, pp. 316322, 2020, [online] Available: https://doi.org/10.18178/ijmlc.2020.10.2.937.

[11] M. Li, B. Wu and Y. Wang, "Comment spam detection via effective features combination", In ICC 2019–2019 IEEE International Conference on Communications (ICC), pp. 1-6, 2019, May.