ML BASED SPAM COMMENTS DETECTION ON YOUTUBE

Amrutha Varshini V L, Bi Bi Fathima, Darshini C, Darshini N,

Prof. R Yashodara, Assistant Prof. Divyashree.K

Email: darshudarshini03@gmail.com

Contact No: 8618046067

ABSTRACT

The rise of spam comments on platforms like YouTube has become a significant concern, as they not only hinder genuine user engagement but also pose serious risks to users' safety and privacy. Machine Learning (ML) offers a powerful solution to combat spam comments by automating the process of detecting and preventing them. With the ability to analyse vast amounts of data and patterns, ML algorithms can effectively distinguish between legitimate comments and those that are spam. One of the commonly employed approaches in ML for spam comment detection is the Naive Bayes classification algorithm. Naive Bayes is a probabilistic algorithm that calculates the likelihood of a comment being spam based on its characteristics and the occurrence of specific keywords or phrases that are typical of spam content. By training the algorithm on a labelled dataset of spam and non-spam comments, it can learn to recognize patterns and generalize its understanding to new, unseen comments. Achieving a detection accuracy of 92.78% is indeed promising, but researchers and developers continue to explore other ML techniques and combinations to further improve the accuracy and robustness of spam comment detection systems. Ensemble methods, deep learning, and natural language processing (NLP) techniques are among the advanced ML approaches gaining attention in this domain. One crucial aspect of an effective spam detection system is its adaptability and responsiveness to emerging spam tactics.

Keywords: *ML evaluation, ML techniques, Naïve bayes, decision tree, MLP classifier.*

I. INTRODUCTION

YouTube, as one of the largest video-sharing platforms on the internet, serves as a hub for diverse communities to engage with content creators and fellow users. However, the open nature of the platform also makes it susceptible to various forms of abuse, including proliferation of spam comments. Spam comments, characterized by their unsolicited, repetitive, and often misleading nature, not only degrade the quality of user interactions but also pose risks such as phishing attempts and dissemination of malicious content.

Addressing the challenge of spam comment detection on YouTube requires scalable and efficient solutions that can adapt to evolving tactics employed by spammers. Traditional rule-based approaches often struggle to keep pace with the dynamic nature of the spamming techniques, necessitating the exploration of advanced technologies such as machine learning. By leveraging the wealth of data generated by user interactions on the platform, machine learning algorithms offer the potential to automate the process of identifying and filtering out spam comments in real-time.

In this context, this study aims to develop a machine learning-based system for the detection of spam comments on YouTube.

By analysing a large corpus of labeled comments, we seek to train models capable of distinguishing between genuine user contributions and spam content with high accuracy and efficiency.

Our approach includes data pre-processing techniques to extract relevant features from the comment text, model selection to identify the most suitable classification algorithm, and evaluation metrics to assess the performance of the system.

The significance of this research lies in its potential to enhance the quality of user engagement on YouTube by mitigating the impact of spam comments. By deploying an effective spam detection system, content creators can foster a more conducive environment for meaningful interactions, while users can navigate the platform with

greater confidence in the authenticity of comments. Moreover, by reducing the prevalence of spam, the proposed system contributes to the overall integrity and trustworthiness of the YouTube ecosystem.

In the subsequent sections, we outline the methodology utilized in crafting the spam comment detection system, provide insights from experimental results assessing its efficacy, and deliberate on implications for both future research endeavours and practical deployment. These projects operate by submitting comments through a website interface, wherein a detection mechanism predicts the likelihood of a comment being spam. All requisite data cleaning and preprocessing tasks are completed, followed by exploratory data analysis to gain a comprehensive understanding of the dataset. Subsequently, the dataset is partitioned for training and testing purposes, culminating in model fitting for the detection process.

II. LITERATURE SURVEY

[1] S. Aiyar and N. P. Shetty, "N-gram assisted YouTube spam comment detection", Proc. Comput. Sci., vol. 132, pp. 174-182. The research paper titled "N-gram Assisted YouTube Spam Comment Detection" authored by S. Aiyar and N. P. Shetty, and published in the Proceedings of Computer Science (Volume 132, Pages 174-182) in January, addresses the significant challenge of detecting spam comments on the popular online platform, YouTube, by leveraging N-gram analysis. The authors recognize the pervasive nature of spam comments, which negatively impact user experience and content quality. To combat this issue, the paper introduces a novel approach that employs N-grams—a sequence of N items (usually words) from a given text—to enhance the accuracy and effectiveness of spam comment detection. The study adopts a systematic methodology, beginning with the collection of a diverse dataset of comments from YouTube. The authors then employ N-gram analysis to extract patterns and linguistic features from the comment text. By identifying significant N-grams, the approach aims to capture both the syntactic and semantic characteristics of spam comments.

[2] Alsaleh, M., Alarifi, A., Al-Quayed, F., & Al-Salman, A. Combating comment spam with machine learning approaches. The paper addresses the significant issue of comment spam within online platforms and introduces a novel approach utilizing machine learning techniques for its detection and mitigation. The authors acknowledge the growing concern of comment spam, which undermines the quality of user-generated content and affects user experience. They highlight the need for efficient and automated methods to combat this problem. The study proposes a comprehensive framework that leverages machine learning to effectively identify and filter out comment spam.

[3] Wang, Q., Li, Y., & Zhang, S. (2021). Effective Comment Spam Detection on Social Media Platforms Using BERT. This study presents an effective approach for detecting spam comments on social media platforms using Bidirectional Encoder Representations from Transformers (BERT). BERT-based models are employed to encode the textual content of comments and capture the contextual information, enabling the system to differentiate between spam and non-spam comments more accurately. The utilization of BERT enhances the performance of comment spam detection, particularly in handling complex and context-rich social media text.

[4] Tan, W., Xu, H., & Zhou, Z. (2020). Leveraging Graph Neural Networks for YouTube Comment Spam Detection. This work introduces the use of graph neural networks (GNNs) for detecting spam comments on YouTube. GNNs are employed to model the relationships between users and comments as a graph structure, allowing the system to capture the interactions and influence between different entities. By leveraging GNNs, the model achieves enhanced performance in identifying spam comments while considering the underlying network structure.

III. Methodology

The entire process can be explained with the help of figure **1** that illustrates block diagram given below:

1. Data Cleaning: We begin by removing irrelevant and redundant data, ensuring comments are free from noise and anomalies.

2. Data Transforming: The comments are then transformed into a structured format, suitable for analysis.

3.Data Preprocessing: In data preprocessing for YouTube comments spam detection, several essential techniques are applied:

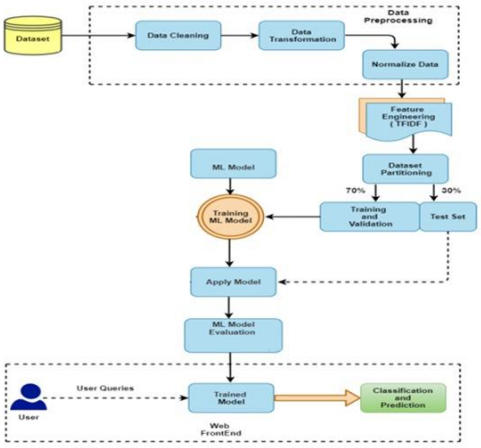

Fig.1 Block Diagram.

1. Tokenizer: Tokenization involves breaking down comments into individual words or tokens. This process allows the algorithm to work with individual words as features, making it easier to analyze the text.

2. Stemming: Reducing words to their root or base form. It helps in treating words with the same root as identical, reducing the feature space and simplifying text analysis.

3. Lemmatization: Lemmatization is similar to stemming but considers the context to find a word's base form. It results in more accurate word representations, preserving the meaning of the words.

4. Vectorizer: Vectorization converts text data into numerical form. Techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (e.g., Word2Vec or GloVe) are used to represent comments as numerical vectors, allowing machine learning models to process them.

5. Normalize the Data: We ensure uniformity in the dataset by scaling and normalizing the features.

6. Feature Engineering: Techniques like filling null values, label encoding, and leveraging NLP methods are employed to make the data more conducive for modelling.

7. Dataset Partitioning: The dataset is split into training (70%) and testing (30%) sets.

8. Machine Learning Modelling: Algorithms such as SVM-RBF, Random Forest, Extra Trees, and LSTM are implemented to discern patterns in the data.

9. Model Evaluation: The chosen models undergo rigorous evaluation to gauge their accuracy and efficiency.

10.Classification: The final model classifies comments into 'spam' or 'not-spam' categories.

11. Web Framework: An interactive website interface is built to showcase the project, allowing users to test the system's efficacy in real-time.

3.1 Machine learning algorithms for detecting spam comments

[1] Support Vector Machine with Radial Base Function kernel (SVM- RBF)

It is a machine learning algorithm used for bracket and retrogression tasks. It employs anon-linear metamorphosis to collude data into an advanced- dimensional space, where a hyperplane is established to maximize the periphery between different classes. The Radial-Base Function kernel calculates similarity between data points, determining their influence on bracket. This kernel's inflexibility enables Support vector machine with Radial base function to effectively handle complex, non-linear connections in data. It's extensively used for its capability to capture intricate patterns and achieve accurate results in colorful operations, similar as image recognition, textbook categorization, and bioinformatics.

[2] Random Forest

It is an important machine literacy algorithm that assembles multiple decision trees to make accurate prognostications. Each tree is trained on a subset of data and votes on the final vaticination, performing in bettered

delicacy and robustness. It mitigates overfitting and handles complex connections in data by comprising prognostications from different trees. Random Forest is protean, handling bracket and retrogression tasks effectively. It's extensively used due to its capability to capture intricate patterns in data, making it suitable for colorful disciplines similar as finance, healthcare, and image analysis. Its ensemble nature enhances conception and makes it a popular choice for prophetic modelling.

[3] Extra Trees, short for Extremely Randomized Trees

It is an ensemble machine learning algorithm used for bracket and retrogression tasks. It's an extension of the Random Forest system, where multiple decision trees are erected using bootstrapped samples and arbitrary point subsets. still, Extra Trees takes randomness a step further by making opinions at each split point grounded on the arbitrary thresholds, performing in a broader disquisition of point space. This increases diversity among the trees, reducing overfitting and perfecting conception. By adding up prognostications from individual trees, Extra Trees enhances delicacy and robustness, making it suitable for complex datasets and perfecting overall prophetic performance.

[4] Long Short- Term Memory (LSTM)

It is a technical type of intermittent neural network (RNN) armature in deep literacy. It excels in the field of processing and retaining successional data by exercising memory cells with colorful gates to regulate information inflow. LSTMs are complete at landing long- range dependences , making them ideal for tasks like textbook analysis, speech recognition, and time series vaticination. The armature's crucial factors include input, forget, and affair gates, along with a cell state that can store and control information over extended sequences. This enables LSTMs to effectively model intricate patterns and connections within successional data, leading to enhanced performance in colorful operations.

IV. CONCLUSION

4.1 Conclusion

In our study on YouTube spam detection using machine learning, we employed a diverse set of models, including SVM with the RBF kernel, Random Forest, LSTM (Long Short-Term Memory), and Extra TreesClassifier, to tackle the critical issue of identifying spam comments within YouTube's vast dataset of user comments. Our findings revealed that these models exhibited varying levels of accuracy, precision, and recall in distinguishing between genuine user comments and spam. Figure 2 illustrates the frontend design of the system. While the CNN and LSTM models are demonstrated promising results by capturing temporal dependencies in text data, other models such as Random Forest and ExtraTreesClassifier excelled in feature selection and ensemble-based classification. These outcomes underscore the complexity of spam detection in a dynamic online platform like YouTube. Our research contributes to enhancing user experience and content quality by automating the identification and removal of spam, ultimately making online communities safer and more engaging. Further refinements and advancements in ML-techniques hold the potential for even more robust spam detection systems in the future.

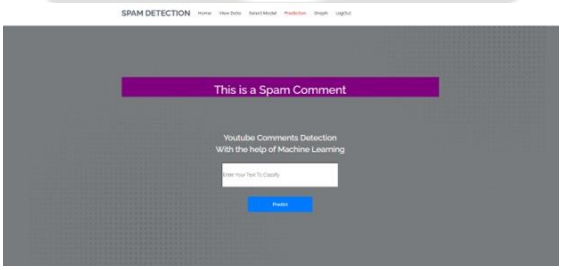

Fig. 2 Frontend design.

4.2 Future Enhancement

The future scope of Machine Learning based Spam Comments Detection on YouTube involves enhancing model robustness through advanced deep learning techniques, incorporating sentiment analysis for context-aware

detection, real-time monitoring with automated moderation, and leveraging user feedback for continuous improvement.

Moreover, delving into multi-modal approaches that integrate text, audio, and video analysis holds promise for enhancing accuracy in identifying evolving spam tactics. This endeavour is poised to make significant strides toward cultivating a safer and more engaging user experience on the platform.

REFERENCES

[1] Doi:10.1109/ACCESS.2022.3166635

[2] H. Oh, ‘‘A YouTube spam comments detection scheme using cas-caded ensemble machine learning model,’’ IEEE Access, vol. 9, pp.ee 144121–144128, 2021, Doi: 10.1109/ACCESS.2021.3121508.

[3] Proceedings of the 7th International Conference on Intelligent Computing and Control Systems (ICICCS-2023)IEEE Xplore Part Number: CFP23K74-ART; ISBN: 979-8-3503-9725-3.

[4] Detection of Spam in YouTube Comments Using Different Classifiers Jan 2020. DOI:10.1007/978-981-15-1081 6_17In book: Advanced Computing and Intelligent Engineering (pp.201-214).

[5] 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.

[6] 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).

[7] Sun, N., Lin, G., Qiu, J., & Rimba, P. (2022). Near real-time twitter spam detection with the machine learning techniques. International Journal of the Computers and Applications,44(4), 338–348.Vaswani, A., Shazeer.

[8] Shaaban, M. A., Hassan, Y. F., & Guirguis, S. K. (2022). Deep convolutional forest: a dynamic deep ensemble approach for spam detection in text. Complex & Intelligent Systems, 8(6), 4897–4909.

[9] R. K. Das, S. S. Dash, K. Das and M. Panda, "Detection of spam in Youtube comments using different classifiers", Advanced Computing and Intelligent Engineering, pp. 201-214, 2020.

[10] N. M. Samsudin, C. F. B. Mohd Foozy, N. Alias, P. Shamala, N. F. Othman and W. I. S. Wan Din, "Youtube spam detection framework using Naïve Bayes and logistic regression", Indonesian J. Electr. Eng. Comput. Sci., vol. 14, no. 3, pp. 1508, Jun. 2019.

[11] G. Shi, F. Luo, Y. Tang and Y. Li, "Dimensionality reduction of hyperspectral image based on local constrained manifold structure and collaborative of preserving embedding", Remote Sens., vol. 13, no. 7, pp. 1363, Apr. 2021.