International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 134-140, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE22
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Automated Spam Detection in YouTube Comments: A Natural Language Processing and Gradient Boosting Approach

[1]B. Dhashvanth Sai, [2]E. Bhargavi, [3]G. Srija Naidu, [4]G. Aditya Srinivas, [5]A. Vimal Kumar

[1,2,3,4,5]B.Tech Student, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India

***Abstract* - The rapid growth of social media platforms such as YouTube, Facebook, Twitter, and TikTok has revolutionized communication but has also led to an increase in spam and harmful content. Detecting spam comments automatically is crucial to maintaining a safe and engaging digital environment. This study proposes a spam detection model using Natural Language Processing (NLP) and XGBoost, a powerful machine learning algorithm known for its high efficiency and predictive accuracy. The model is trained on a dataset containing YouTube comments and utilizes text preprocessing techniques such as tokenization, stopword removal, and lemmatization to enhance detection accuracy. Compared to traditional classifiers like Naïve Bayes and Linear SVM, the proposed NLP-XGBoost model achieves 94% accuracy in classifying spam and non-spam comments. The results demonstrate the potential of machine learning in improving content moderation and safeguarding online interactions.**

Keywords: Spam Detection, NLP, XGBoost, Social Media, Text Classification, Machine Learning, Content Moderation, YouTube Comments.

I. INTRODUCTION

People may now communicate more quickly and easily because to the widespread use of social media platforms like Facebook, Instagram, Twitter, and Tik Tok. People can convey their thoughts, feelings, criticisms, accomplishments, and more through these communication channels. Nevertheless, social networks are frequently used to spread hate speech using demeaning language. The offenses can target a wide range of factors, such as sexual orientation, religion, economic class, ethnicity, and sexism. Therefore, the main issue with this is that the crime is shown to people or groups, which may be detrimental to them. The difficulty in automatically identifying YouTube comments stems from the fact that language used in social networks has a particular format that is inherent to the environment. This context makes use of a number of word abbreviations, as well as intensification and word modification techniques including repeated letters and overuse of punctuation (e.g.: I loved!!!!, what????). As a result, the original text needs to go through a crucial pre-processing step in order to retain its original meaning and maybe align with other postings of a similar nature. In order to protect users' security and mental well-being, software that can identify these kinds of violations in social networks is a crucial improvement. Identifying users who frequently engage in aggressive and cyberbullying behaviour is another application that is being examined. As a result, hostile users need to be blocked and punished. In order to identify YouTube comments in Twitter data, this paper suggests an inquiry. Two methods were selected for this challenge following a survey of the literature: NLP with XGBoost (eXtreme Gradient Boosting).To obtain a format that maintains the original meaning and can in some way align with other posts of a similar nature.

II. LITERATURE REVIEW

M. Bouazizi and T. Ohtsuki (2019) In order to protect users' security and mental well-being, software that can identify these kinds of transgressions in social networks is a crucial improvement. Identifying users who frequently engage in aggressive and cyber bullying behaviour is another application that is being examined. As a result, hostile users need to be blocked and punished. In order to identify YouTube comments in Twitter data, this paper suggests an inquiry. Two methods were selected for this challenge following a survey of the literature: NLP with XGBoost (eXtreme Gradient Boosting).To obtain a format that maintains the original meaning and can in some way align with other posts of a similar nature. The impact of having numerous classes on classification performance is highlighted by the suggested multi-class classification strategy, which obtains an accuracy of 60.2% for seven distinct sentiment classes as opposed to 81.3% for binary classification. However, we suggest a new model to depict the various attitudes and demonstrate how this model aids in comprehending the relationships between them. After that, the model is utilized to examine the difficulties

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 134-140, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE22
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

multi-class classification poses and to indicate potential future improvements in multi-class classification accuracy.

S. Sharma et. al (2019) One of the most significant types of data analytics in the contemporary cyber environment is social media analytics. In order to capitalize on the branding and image-building potential of social media analytics and the detrimental effects of their misuse for defamatory and fraudulent purposes, Understanding the material shared on social media and using thorough analytical

Sanur and Sharma will forecast and assess pertinent data. Anurag, a Jain, talks about social media analytics at the data level, as well as the difficulties and problems that come up when analysing social media data in general. Additionally, a study on Twitter data analysis is presented, in which real-time tweets are gathered and lexicon-based and machine learning approaches are used to comprehend the feelings and thoughts expressed in the text shared on social media. To have a better understanding of social media users' viewpoints, the results effectively categorize their thoughts into three groups: positive, negative, and neutral.

III. ANALYSIS AND DISCUSSION

3.1 Existing System

Naive Bayes utilizes probability computations and presumes the independence of features, rendering it efficient in categorizing text. In contrast, Linear SVM identifies the best hyperplanes to distinct categories and functions effectively in high-dimensional environments. Utilizing Naive Bayes for feature Choosing or pre-processing, succeeded by Linear SVM for classification, we can utilize the advantages of both algorithms and possibly enhance the overall efficiency of the model.

3.2 Existing System Disadvantages

- It has lower predictive accuracy.
- Training a machine requires a significant amount of time.
- The effectiveness of SVM may decrease when handling extremely large datasets

3.3 Proposed System

NLP with XGBoost is a machine learning method that falls under the category of gradient boosting techniques. It is recognized for its rapidity, efficiency, and capability to manage intricate data. XGBoost integrates several weak predictive models, known as decision trees, and trains them in succession to generate precise forecasts. Every following tree is created to rectify the mistakes made by the earlier ones, slowly enhancing the model's effectiveness.

3.4 Proposed System Advantages

- It demonstrates excellent predictive accuracy.
- It is appropriate for extensive datasets.NLP enables computers to comprehend and analyse human language.
- NLP methods help in deriving structured data from unstructured text.
- It has an essential function in machine translation.

3.5 Methodology

Following a literature review, we discovered numerous methods for detecting and classifying YouTube remarks. Deep learning and neural networks have been employed for this aim, and every paperoffered varying outcomes in terms of performance. The solution based on SVM has not emerged, Nonetheless, the creators of the dataset achieved favourable outcomes. The NLP utilizing XGBoost also not discussed in the latest literature, yet it has shown positive outcomes with comparable issues.

3.6 Modules

Data Collection:

This marks the initial genuine move towards the genuine advancement of a machine learning model, gathering information. This is an essential phase that will influence the overall quality of the model; the greater and improved. The more data we obtain, the better our model will function. There are various methods to gather the data, such as web scraping, manual actions, and so on. An NLP-Inspired Approach for Increasing Data to Predict Adverse Events Utilizing Twitter Data collection.

Data:

The dataset includes 24783 rows and 7 columns, which are detailed below.

1. Unnamed: This column lacks a designated name and is typically utilized for reference purposes or identifier for every row of information.

2. Count: This column indicates the frequency of a specific attribute or category's occurrence.

3. YouTube remarks: - This column shows if a tweet or text contains words or expressions that are rude, offensive, or unsuitable.

4. Neither: It signifies a neutral category for content that isn't included in the others.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 134-140, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE22
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

5. Class: It may encompass categories such as "hate speech," "YouTube remarks," or "none" relying on the information presented in the tweet

6. Tweet: It encompasses the material being evaluated for hate speech, YouTube remarks or impartiality

Data Preparation:

Process the data and make it ready for training. Eliminate any items that might need it (get rid of duplicates, fix mistakes, address gaps in data, standardization, and conversion of data types, etc.) Randomizing data eliminates the influence of the specific sequence in which we gathered and/or otherwise arranged our data. Cleaning the dataset's 'tweet' column, to kenizing it, and performing lemmatization.

Additionally utilizing cosine similarity method and fuzzy wuzzy tool to obtain the similarity and matching scores. Divided into training and assessment groups.

Model Selection:

We applied NLP utilizing the XGBoost algorithm. We achieved an accuracy of 94% on the training set, therefore we executed this algorithm.

Analyse and Prediction:

In the current dataset, we selected just 2 characteristics:

1. Class - comprehensive explanation of the data enhancement.

2. Tweet - forecasts whether negative events take place or not.

Accuracy on test set:

We achieved an accuracy of 94% on the testing set.

Saving the Trained Model:

Once you feel sufficiently assured to bring your trained and validated model into production. To prepare the environment, the initial step involves saving it as a .h5 or .pkl file with a library such as pickle. Ensure that you have pickle set up in your environment. Subsequently, Let's import the module and save the model into a .pkl file.

3.7 Technique Used or Algorithm Used

Existing Technique: Linear SVM and Naive Bayes

About deep learning and neural networks in a wider context, the current methodology uses Linear Support Vector Machines (SVM) and Naive Bayes classifiers for a specific goal. A suitable hyperplane to divide classes is searched after by the supervised learning technique known as linear SVM.

On the contrary, Naive Bayes is predicated on the notion of characteristic independence and probabilistic doctrine.

Even though the literature shows mixed outcomes with deep learning and neural networks, the authors highlight the efficiency of Linear SVM and Naive Bayes, stressing their low computational cost and simplicity of use, especially when compared to more sophisticated techniques. As deep learning approaches have been investigated, the authors of the used dataset have shown that Linear SVM and Naive Bayes are appealing alternatives for the given problem due to their ease of use and effectiveness.

3.8 Algorithm Used

Proposed Technique: NLP &XGBoost Classifier

The proposed NLP &XGBoost system combines the power of Natural Language Processing (NLP) techniques with the efficiency and accuracy of XGBoost, an eXtreme Gradient Boosting algorithm. The system begins by preprocessing raw text data, including tasks like tokenization, removing stop words, and applying stemming or lemmatization. Next, relevant features are extracted from the pre-processed text using methods like TF-IDF or word embeddings. XGBoost is then employed to train a classification or regression model on these extracted features and target variables. This integrated approach leverages NLP techniques with the boosting power of XGBoost, offering a robust and accurate solution for various NLP tasks.

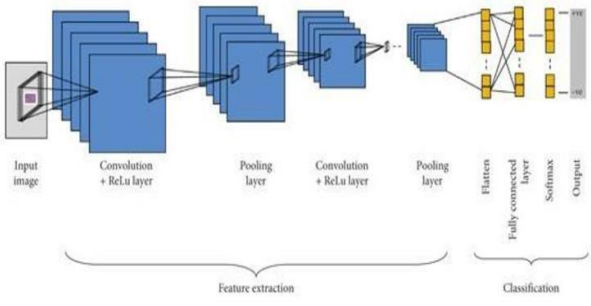


Fig 1: System Architecture

Explanation:

The System Architecture Diagram represents a Convolutional Neural Network (CNN) model used for image classification. The architecture consists of two main stages: Feature Extraction and Classification.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 134-140, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE22
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

1. Feature Extraction

Input Image

The system processes an image through multiple layers.

Convolution + ReLU Layer

Uses filters (kernels) to detect patterns (edges, textures, shapes). ReLU activation adds non-linearity to recognize complex patterns.

Pooling Layer

Retains key features while reducing spatial dimensions. Max/Average pooling improves efficiency and prevents over fitting.

Repeated Convolution + Pooling

Additional convolutional layers extract deeper hierarchical features. Pooling layers further reduce the feature map size.

2. Classification

Uses extracted features for final classification.

Flatten Layer

Converts multidimensional feature maps into a 1D vector forfully connected layers.

Fully Connected Layer

Identifies patterns and relationships between extracted features. Each neuron is connected to all neurons in the previous layer.

SoftMax Layer

Outputs probability values for different classes. Helps in making the final classification decision.

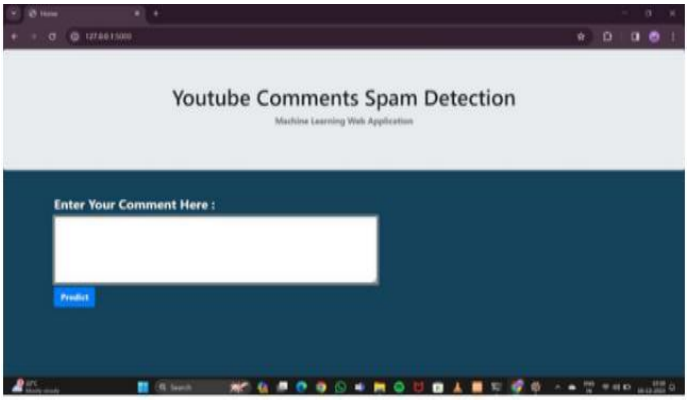
Output Layer

Determines the class of the input image with the final prediction.

IV. RESULT AND DISCUSSION

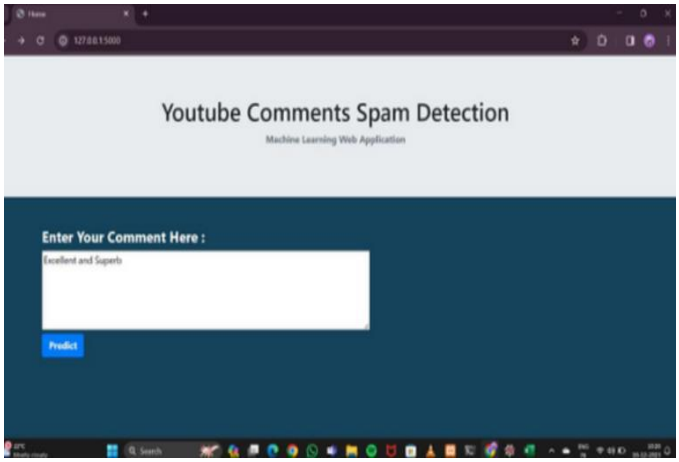
In order to ensure effective data interchange between the client and server, this project was designed utilizing Python with SOCKET and SERVERSOCKET for server-side communication. Socket programming allows for smooth real-time communication, which makes it appropriate for a range of applications that need continuous data transfer. The server process is in charge of processing data, maintaining connections, and responding to client requests. The system's use of sockets guarantees dependable communication while preserving performance and scalability. Cascading Style Sheets (CSS) are utilized in the design aspect to improve the application's user experience and visual attractiveness. A responsive interface and better usability are guaranteed by a well-structured design, which also makes the program easier for users to understand. This architecture offers a strong framework for developing network-based and interactive applications, with room for future improvements like security enhancements, database integration, or sophisticated protocols for communication.

Home Page



The above snapshot indicates home page for users to enter a YouTube comment. A textbox is provided for the user to enter a comment to detect spam comments.

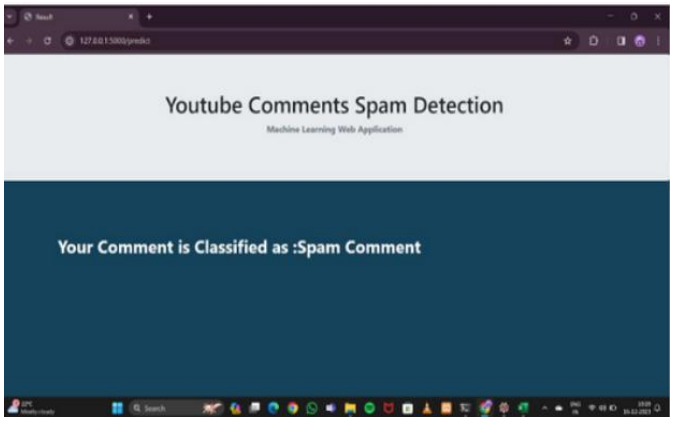
Giving a Spam Comment as Input



The above snapshot indicates home page where a user has entered a YouTube comment. The comment entered is present in the textbox. The user has provided a spam comment as input.

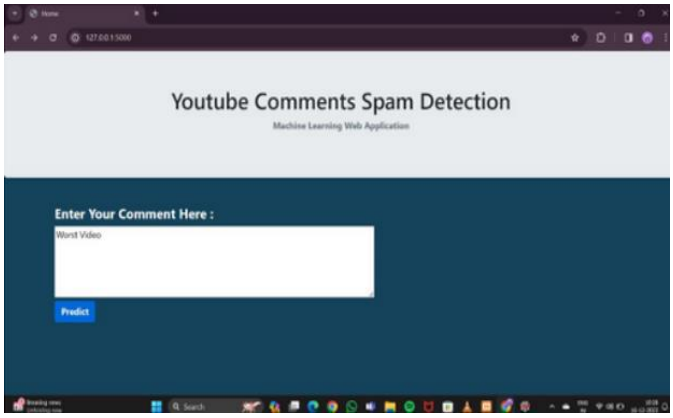
International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 134-140, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE22
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Predicted a Spam Comment



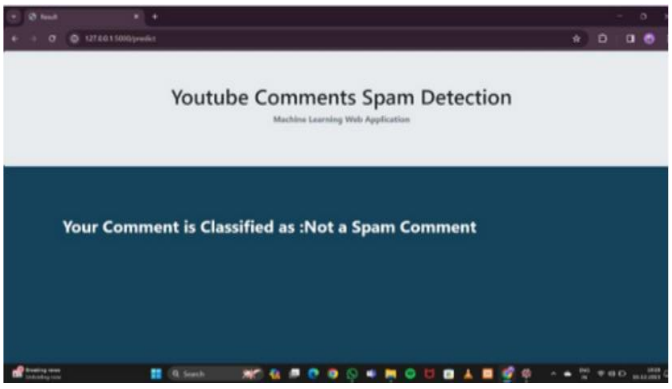
After clicking on the predict button, we will be displayed if the comment is spam or not spam. Here, since a spam comment is provided as input, we will get the display message Your Comment is Classified as: Spam Comment".

Giving a Non-Spam Comment as Input



The above snapshot indicates home page where a user has entered a YouTube comment. The comment entered is present in the textbox. The user has provided a non-spam comment as input.

Predicted a Non-Spam Comment



After clicking on the predict button, we will be displayed if the comment is spam or not spam. Here, since a non-spam comment is provided as input, we will get the display message "Your Comment is Classified as: Not a Spam Comment".

V. CONCLUSION

In this study, we investigated the detection of YouTube comments in tweets using Linear SVM and Naive Bayes classifiers. We discovered throughout our testing phase that the Linear SVM is highly sensitive to the type of data utilized during training. Additionally, it was found that the process of parameter control was hampered by the data standardization using tags. Due of the high standard derivation for the tests with varied seeds, the tests also demonstrated that the evaluation sequence of messages has a significant impact on the classifier's final outcome. If lengthy strings of messages with the same label are provided as input, this is a typical procedure.

REFERENCES

[1] F. Del-Vigna, A. Cimino, F. Dell-Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in First Italian Conference on Cybersecurity, 2017.

[2] M. Bouazizi and T. Ohtsuki, "Multi-class sentiment analysis on twitter: Classification performance and challenges," Big Data Mining and Analytics, vol. 2, no. 3, pp. 181–194, Sep. 2019.

[3] G. Jalaja and C. Kavitha, Sentiment Analysis for Text Extracted from Twitter. Singapore: Springer Singapore, 2019, pp. 693–700.

[4] S. Sharma and A. Jain, "Cyber social media analytics and issues: A pragmatic approach for twitter sentiment analysis," in Advances in Computer Communication and Computational Sciences, S. K. Bhatia, S. Tiwari, K. K. Mishra, and M. C. Trivedi, Eds. Singapore: Springer Singapore, 2019, pp. 473–484.

[5] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Identifying and categorizing youtube comments in social media (offenseval)," arXiv preprint arXiv:1903.08983, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.p. 263– 270.

[7] P. Liu, W. Li, and L. Zou, "Transfer learning for youtube comments detection using bidirectional transformers," in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 87–91.

[8] J. Han, S. Wu, and X. Liu, "Identifying and categorizing youtube comments in social media," in

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 134-140, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE22
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 652–656.

[9] A.Nikolov and V. Radivchev, "Offensive tweet classification with bert and ensembles," in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 691–695.

[10] J. Risch, A. Stoll, M. Ziegele, and R. Krestel, "hpidedis at germeval 2019: Youtube comments identification using a germanbert model," in Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). Erlangen, Germany: German Society for Computational Linguistics & Language Technology, 2019, pp. 403–408. 47.

[11] G. Pitsilis, H. Ramampiaro, and H. Langseth, "Detecting youtube comments in tweets using deep learning," arXiv preprint arXiv:1801.04433, 2018.

[12] Z. Mossie and J.-H. Wang, "Vulnerable community identification using hate speech detection on social media," Information Processing & Management, p. 102087, 2019.

[13] P. Mathur, R. Shah, R. Sawhney, and D. Mahata, "Detecting offensive tweets in hindi-english code-switched language," in Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, 2018, pp. 18–26.

[14] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," arXiv preprint arXiv:1902.09666, 2019.

[15] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," IEEE Access, vol. 6, pp. 13 825–13 835, 2018.

[16] Jahnavi, Y., Kumar, P. N., Anusha, P., & Prasad, M. S. (2022, November). Prediction and Evaluation of Cancer Using Machine Learning Techniques. In International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology (pp. 399-405). Singapore: Springer Nature Singapore.

[17] P. Kadiri, P. Anusha, M. Prabhu, R. Asuncion, V. S. Pavan and J. V. Suman, "Morphed Picture Recognition using Machine Learning Algorithms," 2024 Second International Conference on Advances in Information Technology (ICAIT), Chikkamagaluru, Karnataka, India, 2024, pp. 1-6, doi: 10.1109/ICAIT61638.2024.10690845.

AUTHORS BIOGRAPHY

Budigem Dhashvanth Sai
B.Tech 3rd Year Student, Computer Science and Engineering, Guru Nanak Institute of Technology, Hyderabad, Telangana, India.

Erukula Bhargavi
B.Tech 3rd Year Student, Computer Science and Engineering, Guru Nanak Institute of Technology, Hyderabad, Telangana, India.

Guddati Srija Naidu
B.Tech 3rd Year Student, Computer Science and Engineering, Guru Nanak Institute of Technology, Hyderabad, Telangana, India.

Ganti Aditya Srinivas
B.Tech 3rd Year Student, Computer Science and Engineering, Guru Nanak Institute of Technology, Hyderabad, Telangana, India.

Animalla Vimal Kumar
B.Tech 3rd Year Student, Computer Science and Engineering, Guru Nanak Institute of Technology, Hyderabad, Telangana, India.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 134-140, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE22
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

Citation of this Article:

B. Dhashvanth Sai, E. Bhargavi, G. Srija Naidu, G. Aditya Srinivas, & A. Vimal Kumar. (2025). Automated Spam Detection in YouTube Comments: A Natural Language Processing and Gradient Boosting Approach. In proceeding of International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25), published by *IRJIET*, Volume 9, Special Issue of INSPIRE'25, pp 134-140. Article DOI https://doi.org/10.47001/IRJIET/2025.INSPIRE22

****

