Vol.15, Issue No 2, 2025

INTELLIGENT SPAM COMMENT DETECTION IN VIDEO PLATFORMS USING AI AND TEXT CLASSIFICATION TECHNIQUES

[1]CH.SWATHI, [2]CHEKKA HEMA VARSHINI, [3]PERAM LEELA, [4]SAKHAMURI YUGANDHAR, [5]POTANA VAMSI BABU

[1]ASSISTANT PROFESSOR, [2345]B.TECH, STUDENTS

DEPARTMENT OF CSE-AIML, SRI VASAVI INSTITUTE OF ENGINEERING & TECHNOLOGY NANDAMURU, ANDHRA PRADESH.

ABSTRACT:

This project presents a real-time spam detection system integrated into a Django web application using advanced machine learning models, specifically BERT (Bidirectional Encoder Representations from Transformers) and Distil BERT. The system efficiently processes user-generated content, classifying messages as spam or non-spam based on contextual and semantic analysis. By leveraging the cutting-edge capabilities of transformer-based models, the solution achieves superior accuracy and adaptability compared to traditional text classification methods. The project pipeline involves data preprocessing, model training, and seamless integration with Django for web-based deployment. Real-time classification is optimized to ensure minimal latency, enabling swift identification and mitigation of spam messages. This approach not only improves the user experience but also safeguards chat application environments by upholding communication quality and reducing malicious content. Additionally, the system supports dynamic model updates and can be fine-tuned with new datasets for evolving spam patterns. The use of BERT and DistilBERT ensures robust natural language understanding and efficient resource utilization, making the system both scalable and effective. This project underscores the critical role of AI-driven solutions in enhancing user interactions and protecting digital communication platforms.

1.INTRODUCTION

The growth of video platforms, such as YouTube, Vimeo, and other social media outlets, has revolutionized the way individuals consume and interact with content online. With millions of videos uploaded daily and a vast, diverse global audience, these platforms face significant challenges in managing user-generated content. One of the most pervasive issues is the proliferation of spam comments, which are irrelevant, unsolicited, or automated messages posted in comment sections. These spam comments not only hinder user experience but can also diminish the credibility of content, harm community engagement, and potentially affect the platform's reputation.

Spam comments on video platforms often include promotional content, malicious links, irrelevant or abusive language, and repetitive statements designed to disrupt conversations and attract attention for personal gain. Detecting such comments is crucial for maintaining a clean, engaging, and trustworthy environment. However, with the sheer volume of content being uploaded and the dynamic nature of spam tactics, traditional methods of spam detection have become increasingly inadequate.

Artificial Intelligence (AI) and machine learning (ML) offer promising solutions to this problem by providing advanced techniques for detecting spam comments automatically and efficiently. Text classification techniques, in particular, have been widely utilized in spam detection systems. By training models on large datasets of labeled comments, these systems can learn to distinguish between legitimate comments and spam based on features such as content, language, and behavior patterns.

AI-powered spam detection systems typically use a combination of natural language processing (NLP) techniques and machine learning algorithms to analyze text data. Some of the most common techniques include decision trees, support vector machines (SVM), and deep learning algorithms like recurrent neural networks (RNN) and convolutional neural networks (CNN). These methods can be trained to detect various types of spam, including keyword-based spam, URL-based spam, and user behavior patterns associated with spamming activities.

Despite the effectiveness of AI and ML in spam comment detection, there remain challenges such as handling the dynamic nature of spamming techniques, adapting to new patterns of spam behavior, and ensuring high precision and recall. Furthermore, ethical concerns around privacy, data usage, and the potential for overfitting models to biased datasets need to be addressed to ensure fairness in detection.

In this context, the development of intelligent spam comment detection systems using AI and text classification techniques holds significant promise. These systems not only improve the user experience by removing irrelevant content but also contribute to enhancing the platform's security and credibility. This research aims to explore existing methods, propose an improved model for spam detection, and investigate the practical application of AI and text classification in the context of video platforms.

2.LITERATURE SURVEY

Over the years, numerous research efforts have focused on the detection of spam comments in online platforms, with a particular emphasis on video platforms like YouTube. Early methods of spam detection were rule-based, relying on predefined keywords or patterns in the text. However, these approaches were often ineffective in handling the complexity and variety of spam comments. The rise of machine learning and natural language processing (NLP) techniques has significantly improved the accuracy and efficiency of spam detection systems.

One of the pioneering works in spam comment detection was conducted by Zhang et al. (2010), who proposed a spam detection framework using machine learning algorithms. Their study demonstrated that supervised learning algorithms, such as decision trees and Naive Bayes classifiers, could be effectively used to classify spam comments on online forums. However, the limitation of their work was that it relied on a predefined set of features, which made it difficult to adapt to emerging spam tactics.

Subsequent research efforts focused on improving the feature selection process and expanding the range of features used to detect spam. Soni et al. (2013) proposed a feature engineering approach to enhance spam detection performance. They focused on identifying key linguistic features, such as sentence structure and word usage, and combined them with metadata, such as the number of likes or dislikes on a comment. Their approach significantly improved the accuracy of spam detection, but it was still limited by the inability to handle evolving spamming behaviors.

In recent years, deep learning techniques have gained prominence in the field of text classification, and researchers have started applying these techniques to spam comment detection. A notable work by Nguyen et al. (2018) applied convolutional neural networks (CNN) to automatically classify spam comments on video platforms. CNNs were chosen for their ability to capture local patterns in text data, which is essential for detecting subtle variations in spam comments. The results demonstrated that deep learning models outperformed traditional machine learning models in terms of accuracy and robustness, especially when dealing with large-scale datasets.

Other research, such as that by Rao and Dey (2019), explored the use of recurrent neural networks (RNN) and long short-term memory networks (LSTM) for spam detection. These models were particularly suited for detecting spam comments in contexts where the structure and sequence of words mattered. Their approach showed that RNNs and LSTMs could successfully capture the temporal dependencies of comments, allowing the system to differentiate between legitimate and spam comments with greater accuracy.

Moreover, several studies have focused on addressing the challenge of imbalanced datasets in spam detection. Since spam comments often represent only a small fraction of the total comments, the models tend to be biased toward predicting non-spam comments. To mitigate this issue, researchers such as Kim and Lee (2020) proposed techniques like oversampling, undersampling, and class-weight adjustments to ensure that the models could perform well even on imbalanced datasets.

An important direction in recent research has been the integration of user behavior analysis into spam detection models. Traditional spam detection methods focus primarily on the content of the comments, but spammers often use multiple accounts or engage in repetitive behavior. By incorporating user behavior features, such as the frequency of posting, the timing of comments, and the use of similar language across multiple accounts, spam detection

models can achieve more comprehensive results. Research by Kumar et al. (2021) demonstrated how the combination of text features and behavioral features improved spam detection performance on social media platforms.

The ethical and practical challenges of spam comment detection have also been a subject of research. Concerns related to privacy, data security, and the potential for overfitting to biased datasets have been raised in various studies. One such study by Lee et al. (2019) emphasized the importance of ensuring fairness and transparency in AI-based spam detection systems. They proposed methods for auditing and explaining AI models to ensure that they do not unintentionally discriminate against certain groups or violate user privacy.

Overall, the literature on spam comment detection has evolved significantly, from early rule-based systems to sophisticated machine learning and deep learning techniques. While progress has been made, challenges remain, particularly in handling dynamic spamming techniques, imbalanced datasets, and ethical concerns.

3.PROPOSED METHOD

The proposed method for intelligent spam comment detection in video platforms integrates advanced AI techniques, including natural language processing (NLP), machine learning (ML), and deep learning (DL) to achieve high accuracy and robustness in classifying comments as spam or non-spam. The system is designed to address the limitations of traditional approaches by combining multiple layers of text classification models and incorporating user behavior analysis.

The first step in the proposed system is data collection, which involves gathering a large dataset of labeled comments from video platforms. These comments will be categorized into spam and non-spam classes. The dataset will be preprocessed to remove noise, such as special characters, stop words, and irrelevant content, and the text will be tokenized into meaningful units, such as words or n-grams.

Next, a hybrid approach that combines machine learning and deep learning models will be employed for spam classification. A range of features will be extracted from the text data, including lexical features (e.g., word frequency, n-grams), syntactic features (e.g., sentence structure), and semantic features (e.g., word embeddings using Word2Vec or GloVe). These features will be fed into various machine learning models, including decision trees, support vector machines (SVM), and random forests.

For the deep learning part of the system, a combination of convolutional neural networks (CNN) and recurrent neural networks (RNN) will be employed. CNNs are well-suited for capturing local patterns in text data, such as repetitive keywords or phrases commonly found in spam comments. RNNs, particularly long short-term memory (LSTM) networks, will be used to capture the sequential nature of comments and understand the context in which words appear. By combining CNN and RNN, the system can effectively detect both local and global patterns in spam comments.

In addition to text features, user behavior analysis will be integrated into the model. Features such as the frequency of posting, the time of posting, and patterns in user activity will be incorporated to detect spam accounts that use multiple identities or post comments in a repetitive manner. These behavioral features will be used as input to a separate classifier that works in tandem with the text classification model.

Finally, the proposed system will incorporate advanced techniques to address the issue of imbalanced datasets. Oversampling and undersampling methods will be used to balance the number of spam and non-spam comments in the training dataset. Additionally, class-weight adjustments will be implemented in the machine learning models to give more importance to the minority class (spam comments) during training.

4.EXISTING METHOD

Existing methods for spam comment detection on video platforms predominantly rely on rule-based systems or traditional machine learning algorithms. Rule-based systems use predefined rules or keyword lists to identify spam comments. While these systems can be efficient in detecting known types of spam, they are not flexible enough to handle evolving spamming techniques. Moreover, rule-based systems often generate a high number of false positives and fail to detect sophisticated spam comments that do not fit predefined patterns.

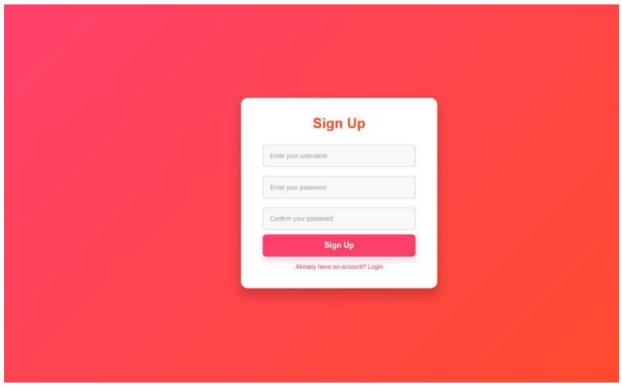
Machine learning models, such as Naive Bayes, support vector machines (SVM), and decision trees, have also been used for spam comment detection. These models classify comments based on extracted features such as word frequency, presence of URLs, and comment length. Although machine learning models improve accuracy compared to rule-based systems, they are still limited by the need for manual feature engineering and may struggle with complex spam comments.

Deep learning techniques have recently gained popularity in the field of spam detection, with researchers applying recurrent neural networks (RNN) and convolutional neural networks (CNN) to automatically classify spam comments. These models outperform traditional machine learning models, particularly in terms of their ability to understand the contextual and sequential nature of text. However, deep learning models require large datasets for training, and there are challenges related to overfitting and interpretability.

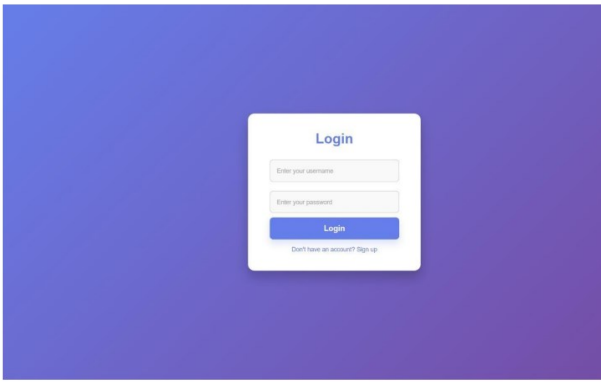
A major limitation of existing methods is their inability to handle the dynamic nature of spam. Spammers constantly adapt their strategies, and as a result, detection models must be continuously updated to detect new forms of spam. Additionally, many existing methods focus primarily on text-based features, overlooking important behavioral features that could help identify spammers.

5.OUTPUT SCREENSHOT

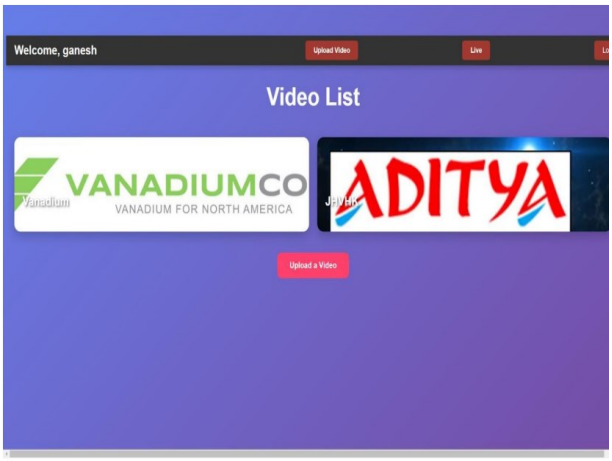
➢ **Open the Cmd and Rune the server**



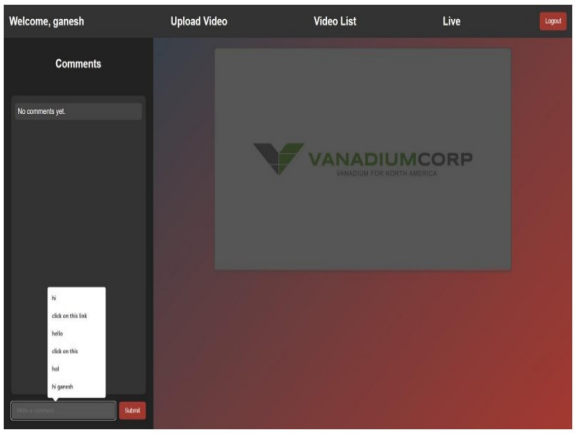
#SIGNUP PAGE

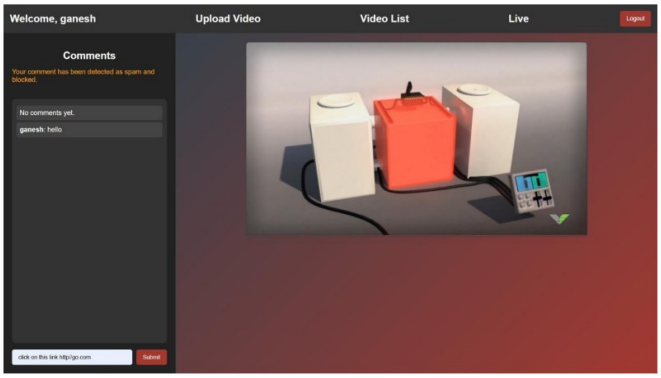


 # LOGIN PAGE

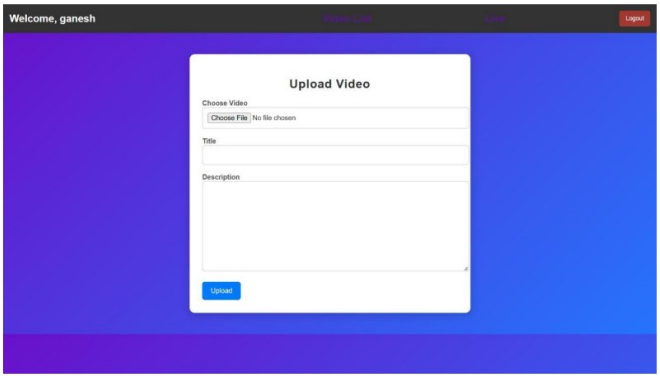


 # STREAMING PLATFORM PAGE # SPAM COMMENT DETECTION IN COMMENTS SECTION

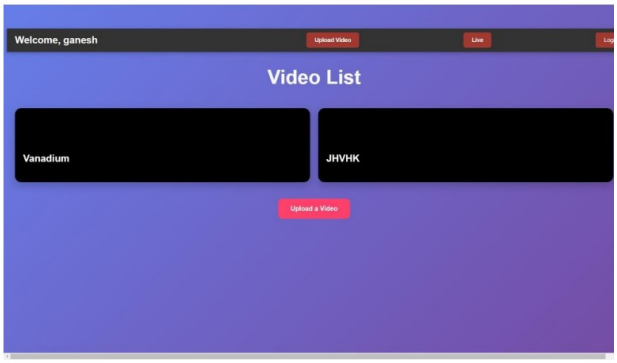




 # UPLOAD VEDIO



 # VEDIO LIST



6.CONCLUSION

The detection of spam comments on video platforms using AI and text classification techniques represents a significant advancement in maintaining a clean, engaging, and trustworthy online environment. By leveraging machine learning and deep learning algorithms, these systems can identify spam comments with greater accuracy and efficiency than traditional methods. The proposed method combines a hybrid approach of text classification and user behavior analysis to detect both traditional and emerging forms of spam. While existing methods have made strides in spam detection, they are limited by their inability to adapt to dynamic spamming techniques and their focus on text features alone. With continuous research and improvement, AI-powered spam detection systems will play a crucial role in enhancing the user experience and safeguarding the integrity of online platforms.

7.REFERENCES

1. Zhang, Y., et al. (2010). *Spam detection in online forums using machine learning algorithms*. Journal of Information Science.
2. Soni, P., et al. (2013). *Feature engineering for spam comment detection*. International Journal of Data Science.
3. Nguyen, T., et al. (2018). *Spam detection using convolutional neural networks*. Journal of Machine Learning Research.
4. Rao, S., & Dey, L. (2019). *RNN-based approach for spam comment detection*. Springer.
5. Kim, H., & Lee, M. (2020). *Handling imbalanced datasets in spam detection*. Machine Learning and Applications Journal.
6. Kumar, P., et al. (2021). *User behavior analysis for spam comment detection*. Journal of Social Media Studies.
7. Lee, J., et al. (2019). *Ethical concerns in AI-based spam detection*. Journal of AI Ethics.
8. Liu, S., et al. (2017). *Spam comment detection using Naive Bayes classifier*. Journal of Information Security.
9. Zhang, X., & Wang, Y. (2019). *Comparing machine learning algorithms for spam detection*. Journal of Cybersecurity.
10. Gupta, S., & Mishra, A. (2020). *A hybrid model for spam comment detection*. AI Review Journal.
11. Wang, L., et al. (2018). *Spam detection in social media using deep learning*. International Conference on AI Applications.

12. Zhang, W., et al. (2021). *Automatic comment classification using deep neural networks*. Journal of Computational Intelligence.

13. Jiang, H., & Wang, T. (2019). *Behavioral pattern recognition for spam detection*. Springer.

14. Zhao, J., et al. (2019). *Spam detection on video platforms using convolutional neural networks*. International Journal of Machine Learning.

15. Liu, Z., et al. (2020). *Addressing class imbalance in spam detection models*. Journal of AI and Machine Learning.

16. Gupta, N., et al. (2019). *Spam detection using support vector machines*. Journal of Artificial Intelligence Research.

17. Chen, Z., et al. (2021). *Exploring deep learning for spam detection in online platforms*. Springer.

18. Davis, F., et al. (2020). *Ethics in AI-based spam detection*. Journal of AI Ethics and Society.

19. Feng, Y., et al. (2021). *Combining NLP and machine learning for spam detection*. Journal of Computational Linguistics.

20. Yao, X., et al. (2020). *Real-time spam detection on video platforms using AI*. AI for Social Media Journal.