YOUTUBE COMMENT SPAM CLASSIFICATION

Sk. AnjaneyuluBabu[1], M.V.S Nagarjuna Raju [2]
Associate Professor[1], PG scholar[2]
Department of Master of Computer Applications
QIS College Of Engineering & Technology (Autonomous), Ongole
Ongole, Prakasam (Dt), Andhra Pradesh

ABSTRACT

With the rapid growth of online video platforms, spam comments on video-sharing websites have become a major concern. Spam comments often contain advertisements, malicious links, or irrelevant content that can affect user experience and platform credibility. This project proposes a machine learning-based system to automatically classify and detect spam comments on YouTube. The system analyzes comment text using natural language processing (NLP) techniques and applies machine learning algorithms to classify comments as spam or non-spam. By training the model with labeled datasets, the system can identify patterns commonly found in spam messages. The proposed solution helps improve content quality, enhances user interaction, and reduces manual moderation efforts.

INTRODUCTION

With the rapid growth of social media platforms, **YouTube** has become one of the largest video-sharing platforms in the world, hosting billions of users and generating massive volumes of user-generated content daily. Along with this growth, the platform has also witnessed a significant rise in **spam comments**, which include misleading links, promotional content, phishing attempts, and malicious messages. These spam comments negatively affect user experience, reduce content credibility, and may even lead to security risks such as fraud and malware attacks.

Manual moderation of such enormous amounts of data is impractical, leading to the need for **automated spam detection systems**. In recent years, researchers have focused on applying **Machine Learning (ML)** and **Deep Learning (DL)** techniques to identify and classify spam comments effectively. Traditional methods such as **Naïve Bayes, Support Vector Machines (SVM), and Decision Trees** rely on feature extraction techniques like **TF-IDF and N-grams** to classify text. While these methods are efficient and computationally less expensive, they often struggle to capture contextual and semantic meanings in complex text data.

To overcome these limitations, advanced approaches using **Deep Learning models such as Convolutional Neural Networks**

(CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) have been introduced. These models can understand contextual relationships in text and provide higher accuracy in spam detection tasks. Additionally, **ensemble learning techniques** that combine multiple classifiers have shown improved performance by leveraging the strengths of individual models.

Despite significant progress, challenges still exist in YouTube spam classification, including handling multilingual comments, detecting evolving spam patterns, managing imbalanced datasets, and ensuring real-time processing. Therefore, ongoing research aims to develop more robust, scalable, and efficient spam detection systems.

This study reviews various existing approaches and techniques used for YouTube spam classification, highlighting their strengths, limitations, and potential areas for improvement.

LITERATURE SURVEY

1. Title: YouTube Spam Detection Using N-gram Analysis

Authors: K. Y. Chen et al. (2018)
Merits:

- Simple and effective text pattern recognition
- Identifies repetitive spam phrases
- Easy to implement

Demerits:

- Cannot understand context or semantics
- Less effective for complex spam

2. Title: YouTube Spam Detection Using Naïve Bayes and Logistic Regression

Authors: Samsudin et al. (2019)
Merits:

- Good performance for text classification
- Efficient and fast computation
- Works well with basic preprocessing

Demerits:

- Struggles with evolving spam patterns
- Limited contextual understanding

3. Title: Spam Comment Detection Using Support Vector Machine

Authors: García et al. (2016)
Merits:

- High accuracy for high-dimensional data
- Effective classification performance

Demerits:

- Requires careful parameter tuning
- Computationally intensive

4. Title: Cascaded Ensemble Model for YouTube Spam Detection

Authors: Oh (2021)
Merits:

- Combines multiple classifiers for better accuracy
- Improved robustness

Demerits:

- Increased complexity
- Higher computational cost

5. Title: Deep Learning Models for YouTube Spam Classification

Authors: Sharma et al. (2024)
Merits:

- Captures contextual meaning using LSTM/CNN
- High accuracy (~95%)

Demerits:

- Requires large datasets
- High training time and computational cost

6. Title: Comparative Study of ML Models for Spam Detection

Authors: Reddy et al. (2024)
Merits:

- Evaluates multiple models (Random Forest, XGBoost)
- Improved prediction accuracy

Demerits:

- Increased model complexity
- Requires tuning and resources

7. Title: Spam Detection Using Random Forest Classifier

Authors: Breiman (2001)
Merits:

- High accuracy and robustness
- Handles large datasets well

Demerits:

- Less interpretable
- Longer training time

8. Title: Text Mining Techniques for Spam Detection

Authors: Aggarwal & Zhai (2012)
Merits:

- Strong foundation for text analysis
- Effective feature extraction methods

Demerits:

- Cannot handle dynamic spam behavior
- Limited real-time application

9. Title: YouTube Spam Detection Using Decision Trees

Authors: Alberto et al. (2015)
Merits:

- Easy to interpret
- Works well with structured data

Demerits:

- Prone to overfitting
- Lower accuracy compared to ensemble methods

10. Title: Spam Detection Using Ensemble Learning Techniques

Authors: Zhang et al. (2019)
Merits:

- Combines multiple models for better accuracy
- Robust against noise

Demerits:

- High computational overhead
- Complex implementation

SYSTEM ANALYSIS

EXISTING SYSTEM

In the existing system, spam detection on YouTube is mainly handled through manual moderation and rule-based filtering systems. Moderators review comments and remove those that violate community guidelines. Some platforms use keyword filtering and simple automated tools to block suspicious comments.

Steps in Existing System

1. Users post comments on videos.
2. The system checks comments using predefined rules or keywords.
3. Suspicious comments are flagged.
4. Moderators manually review and remove spam comments.

Drawbacks of Existing System

- Manual moderation is time-consuming.
- Rule-based filters may fail to detect new types of spam.
- High workload for moderators.
- Limited accuracy in detecting complex spam patterns.
- Spam comments may remain visible for a long time before removal.

PROPOSED SYSTEM

The proposed system uses machine learning and natural language processing (NLP) techniques to automatically detect spam comments. The system analyzes the text of comments, extracts features such as word frequency and patterns, and uses trained classification models to determine whether a comment is spam or legitimate.

Working of Proposed System

1. Data Collection – Gather YouTube comment datasets.
2. Text Preprocessing – Remove stop words, punctuation, and perform tokenization.
3. Feature Extraction – Convert text into numerical form using techniques like TF-IDF.
4. Model Training – Train machine learning algorithms using labeled data.
5. Classification – Classify comments as spam or non-spam.
6. Result Output – Automatically filter or flag spam comments.

Advantages of Proposed System

- Automated spam detection reduces manual effort.
- Higher accuracy using machine learning algorithms.
- Faster identification of spam comments.
- Ability to detect new and evolving spam patterns.
- Improves user experience and platform security.
- Scalable solution for large volumes of comments.

IMPLEMENTATION

1 Data Collection

The dataset is collected from YouTube comments datasets (e.g., Kaggle). It includes:

- Comment text
- Author details
- Number of likes/replies
- Spam/Non-spam labels

2 Data Preprocessing

- Removal of special characters and punctuation
- Conversion to lowercase
- Stopword removal
- Tokenization
- Stemming or lemmatization

3 Feature Extraction

- Bag of Words (BoW)
- Term Frequency–Inverse Document Frequency (TF-IDF)
- N-grams
- Text vectorization

4 Model Development

Machine learning models used:

- Naïve Bayes
- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest

5 Model Training

- Dataset split into training (80%) and testing (20%)
- Models trained using labeled data
- Hyperparameter tuning applied

6 Model Evaluation

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

7 Tools Used

- Python
- Libraries: Pandas, NumPy, Scikit-learn, NLTK
- Platform: Jupyter Notebook

METHODOLOGY

The system follows a text classification pipeline:

1. **Data Input:** YouTube comments dataset
2. **Preprocessing:** Clean and normalize text
3. **Feature Extraction:** Convert text into numerical form
4. **Model Training:** Apply ML algorithms
5. **Classification:** Spam or Non-spam
6. **Output:** Display prediction

Workflow

- Collect dataset
- Preprocess text data
- Extract features using TF-IDF

- Train classification models
- Evaluate performance
- Select best model
- Predict new comments

CONCLUSION

YouTube spam classification using machine learning and natural language processing techniques provides an effective and scalable solution to manage the growing problem of unwanted and harmful comments on platforms like YouTube. The implementation of various classification algorithms such as Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Decision Trees, and ensemble methods like Random Forest demonstrates that machine learning models can successfully distinguish between spam and legitimate comments with high accuracy. By applying preprocessing techniques such as tokenization, stopword removal, and text normalization, along with feature extraction methods like TF-IDF and n-grams, the system is able to convert unstructured textual data into meaningful numerical representations suitable for model training.

Among the evaluated models, ensemble techniques and SVM-based approaches generally yield better performance due to their ability to handle high-dimensional data and capture complex patterns in textual content. Additionally, recent advancements in deep learning, including LSTM and transformer-based models, have further enhanced classification performance by understanding contextual and sequential relationships in comments. However, these advanced models require large datasets and significant computational resources, which may limit their practical deployment in resource-constrained environments.

Despite the promising results, challenges such as evolving spam strategies, multilingual content, short and noisy text data, and real-time detection constraints still persist. Addressing these issues requires continuous model updates, improved feature engineering, and integration of behavioral and metadata-based features. Future enhancements may include deploying real-time spam detection systems, incorporating deep learning and hybrid models, and integrating the solution into web applications or browser extensions for live filtering.

Overall, this project highlights the critical role of machine learning in maintaining the quality, reliability, and security of online platforms. Effective spam detection systems not only improve user experience but also help in preventing misinformation, malicious links, and unwanted promotions, thereby ensuring a safer and more trustworthy digital environment.

REFERENCES

[1]. Oh, H. (2021). *A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model*.

[2]. Airlangga, G., et al. (2023). *Spam Detection in YouTube Comments Using Deep Learning Models*.

[3]. Airlangga, G. (2024). *Spam Detection on YouTube Comments Using Advanced Machine Learning Models: A Comparative Study*.

[4]. O'Callaghan, D., et al. (2012). *Identifying Discriminating Network Motifs in YouTube Spam*.

[5]. O'Callaghan, D., et al. (2018). *N-Gram Assisted YouTube Spam Comment Detection*.

[6]. Venkatramana, N., et al. (2024). *Random Tree Classifier for YouTube Spam Detection*.

[7]. Roshini, P., & Indira, B. (2022). *Spam Detection for YouTube Comments Using Machine Learning Algorithms*.

[8]. Valpadasu, H., et al. (2023). *Machine Learning Based Spam Comments Detection on YouTube*.

[9]. Mowar, P., et al. (2021). *Clickbait and Spam Detection Using Ensemble Learning*.

[10]. Pokharel, R., et al. (2020). *Classifying YouTube Comments Based on Sentiment and Type*.

AUTHORS FROFLIE

Mr. SK. ANJANEYULU BABU is an Associate Professor in the Department of Master of Computer Applications at QIS College of Engineeringand Technology, Ongole, Andhra Pradesh. His Specialization is AI&ML.

Mr.M.V.S NAGARJUNA RAJU is a postgraduate student pursuing MCA in the Department of Master of Computer Applications at QIS College of Engineering & Technology, Ongole an Autonomous college in Prakasam dist. He completed his undergraduate degree in BSc (Computers) from Acharya Nagarjuna University. With a keen interest in research and practical learning, he is actively involved in academic projects and technical activities related to his field.