# Classification of YouTube Spam Comments Using Machine Learning and Deep Learning Techniques

Submitted by :
Group no.: 2

| Name | Roll no |
|---|---|
| Girish S | AM.EN.U4AIE22044 |
| Anuvind M P | AM.EN.U4AIE22010 |
| Harishankar Binu Nair | AM.EN.U4AIE22023 |
| Thazhai Mugunthan G | AM.EN.U4AIE22051 |

# ABSTRACT

## *Informal description*

Sometimes, when you look at YouTube comments, you see spam messages with weird links or repetitive phrases. These are often posted by spam bots, and the goal is to find a way to automatically detect and remove them. We need to develop a system that can identify these suspicious comments and differentiate them from real ones.

To achieve this, we can use techniques like analyzing the content of the comments, observing posting patterns, and identifying common spam traits. This helps clean up the comments section, making it more useful and safer for everyone.

## *Formal description*

YouTube, as one of the largest video-sharing platforms, hosts a vast amount of user-generated content, including comments. These comments play a significant role in user engagement and community building. However, the presence of spam comments can disrupt user experience, diminish the quality of discussions, and potentially spread harmful content. Therefore, it is essential to develop effective methods to automatically detect and filter out spam comments on YouTube.

The primary objective of this project is to evaluate and compare the effectiveness of various machine learning and deep learning models in classifying YouTube comments as spam or not spam. The models under consideration include:

- Naive Bayes
- Recurrent Neural Networks (RNN)
- Transformers (BERT)
- GPT

# 1.INTRODUCTION

## *Motivation :*

Spam bots in YouTube comments clutter the section, making it challenging for users to find meaningful discussions and genuine feedback. Solving this problem will create a cleaner, more engaging environment for users. Furthermore, reducing spam helps protect users from potentially harmful links and scams that can be found in these comments. This enhances the overall safety of the platform, leading to a more enjoyable and secure experience for all YouTube users.

## *Benefits of the solution :*

1. **Improved User Experience**: Enhances interactions by removing irrelevant or harmful comments.
2. **Enhanced Content Quality**: Promotes genuine discussions, making valuable feedback more visible.
3. **Increased User Trust**: Builds trust by effectively managing spam, encouraging user participation.
4. **Protection Against Harmful Content**: Shields users from malicious links and misleading information.
5. **Support for Content Creators**: Allows creators to focus on real feedback, improving content and community.
6. **Reduced Moderation Costs**: Lowers the need for manual moderation, saving resources.
7. **Scalability**: Efficiently handles large volumes of comments, ensuring consistent spam management.

## *Use of Solution:*

Identification of the most effective model for classifying spam comments on YouTube. Insights into the trade-offs between model complexity, training time, and classification performance.

Recommendations for implementing the most suitable model in real-world spam detection systems on YouTube.
Effectively detecting spam comments on YouTube is vital for maintaining the quality and integrity of discussions on the platform. By comparing the performance of various machine learning and deep learning models, this project aims to identify the most suitable approach for automatic spam detection. The findings will provide valuable insights and practical recommendations for deploying spam classification systems on YouTube, enhancing user experience and engagement.

# 2.Dataset Finalization

### Dataset 1 : 5000 Youtube Spam/Not Spam dataset (link) :

This dataset contains 5000 YouTube comments along with associated metadata. It includes information about the commenter, the content of the comment, the timestamp, the number of likes and replies, and a label indicating whether the comment is spam or not.

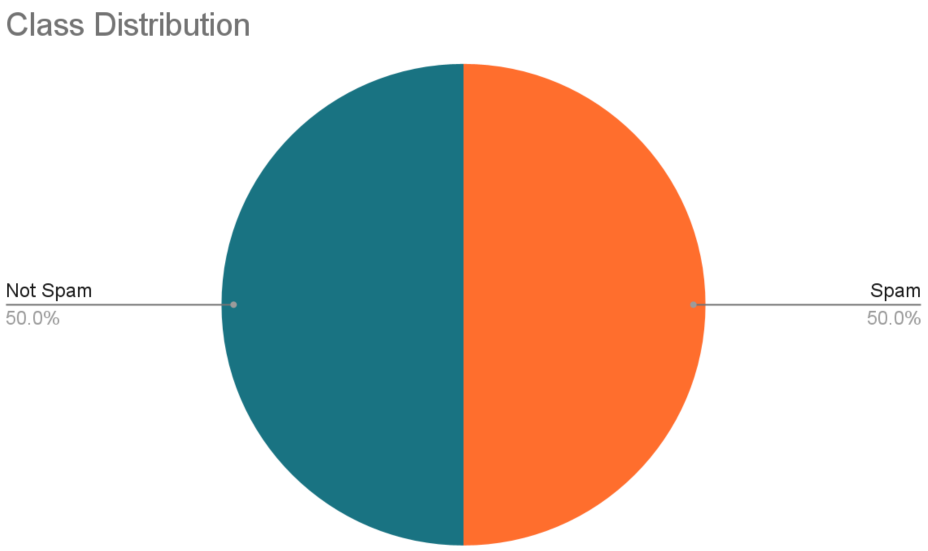
**Rows:** 5000

**Columns:** 6

### Features Description

1. **Name, Time:**
    - **Importance:** Irrelevant for the model.
2. **Comment:**
    - **Type:** String
    - **Description:** The text of the YouTube comment.
    - **Importance:** The primary feature used to determine if a comment is spam. Analyzing the text content helps in identifying spam patterns or phrases.

3. **Likes:**
    - **Type:** Integer
    - **Description:** The number of likes the comment received.
    - **Importance:** Can indicate engagement, but spam comments can sometimes have artificially inflated "like" counts. Not used in the model.
4. **Reply Count:**
    - **Type:** Integer
    - **Description:** The number of replies to the comment.
    - **Importance:** Can show engagement level, but not used in the model.
5. **Spam:**
    - **Type:** Integer (0 or 1)
    - **Description:** A label indicating whether the comment is spam (1) or not (0).
    - **Importance:** This label is crucial for supervised learning, helping the model to learn the difference between spam and non-spam comments.



## Dataset 2 : Youtube Spam Collection (link)

**Rows**: 1956

**Columns**: 5

**Features Description:**

**1. COMMENT_ID**

- **Type**: String
- **Description**: A unique identifier for the comment.
- **Importance**: Irrelevant for the model.

**2. AUTHOR**

- **Type**: String
- **Description**: The name of the commenter.
- **Importance**: Irrelevant for the model.

**3. DATE**

- **Type**: String/Datetime
- **Description**: The timestamp when the comment was made.
- **Importance**: Irrelevant for the model.

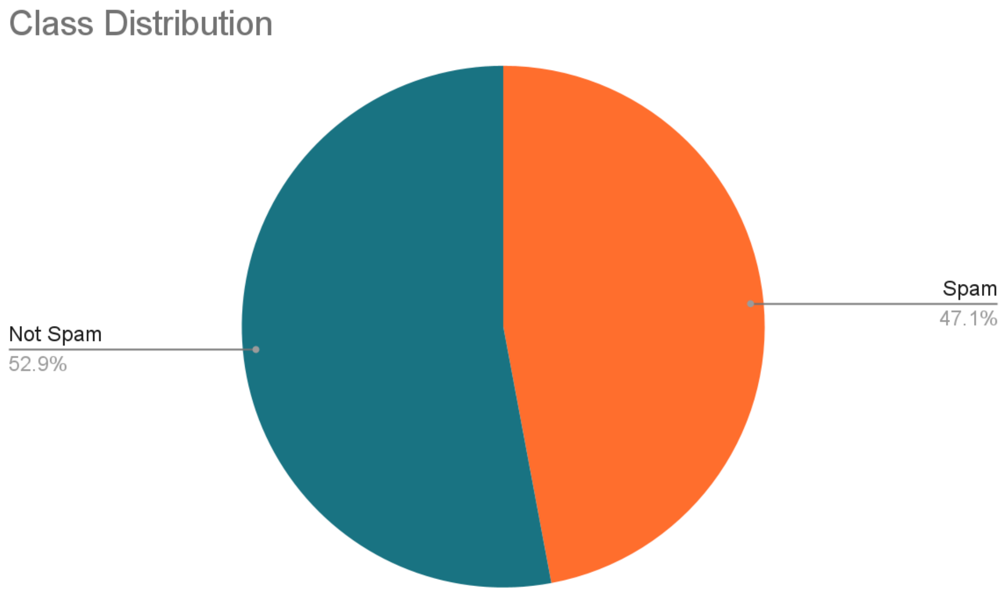
**4. CONTENT**

- **Type**: String
- **Description**: The text of the YouTube comment.
- **Importance**: The primary feature used to determine if a comment is spam. Analyzing the text content helps in identifying spam patterns or phrases.

**5. CLASS**

- **Type**: Integer (0 or 1)
- **Description**: A label indicating whether the comment is spam (1) or not (0).

- **Importance**: This label is crucial for supervised learning, helping the model to learn the difference between spam and non-spam comments.



**Dataset 3 : Youtube Spam Merged Data [(link)]**

**Total Rows:** 1956

**Total Columns:** 7

**Features Description:**

**1. Unnamed: 0**

- **Type:** Integer
- **Description:** Index column.
- **Importance:** Irrelevant for the model.

**2. Unnamed: 1**

- **Type:** Integer
- **Description:** Secondary index column.
- **Importance:** Irrelevant for the model.

**3. COMMENT_ID**

- **Type:** String
- **Description:** A unique identifier for the comment.
- **Importance:** Irrelevant for the model.

**4. AUTHOR**

- **Type:** String
- **Description:** The name of the commenter.
- **Importance:** Irrelevant for the model.

**5. DATE**

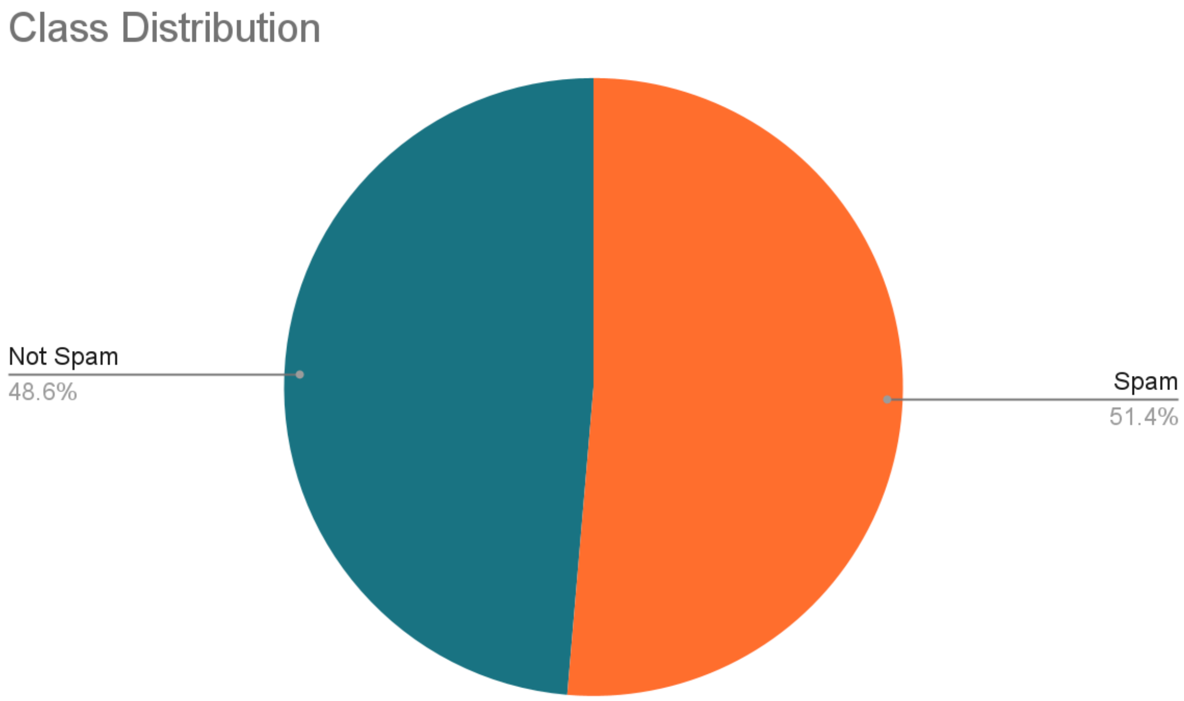
- **Type:** String/Datetime
- **Description:** The timestamp when the comment was made.
- **Importance:** Irrelevant for the model.

**6. CONTENT**

- **Type:** String
- **Description:** The text of the YouTube comment.
- **Importance:** The primary feature used to determine if a comment is spam. Analyzing the text content helps in identifying spam patterns or phrases.

**7. CLASS**

- **Type:** Integer (0 or 1)
- **Description:** A label indicating whether the comment is spam (1) or not (0).
- **Importance:** This label is crucial for supervised learning, helping the model to learn the difference between spam and non-spam comments.


Class Distribution
Not Spam
48.6%
Spam
51.4%