**22AIE213**

**Introduction to Machine Learning**

# **YouTube Spam Comment Classification - A Review**

## **Abstract**

**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/NotSpam.**

## **Introduction**

YouTube, as one of the largest video-sharing platforms, receives millions of comments daily. While user engagement is beneficial, the platform is also plagued by spam comments that degrade the user experience. These spam comments often include misleading links, advertisements, and irrelevant content, disrupting meaningful interactions and potentially leading users to malicious sites. The sheer volume of comments makes it impractical for manual moderation, necessitating the development of automated spam detection systems. Effective spam detection not only enhances user experience but also maintains the integrity and safety of the platform. As the techniques employed by spammers evolve, it becomes crucial to continuously improve and adapt spam detection methods to stay ahead of these disruptions.

## **Objectives**

The primary objective of this project is to develop a robust spam detection system for YouTube comments using different machine learning models. The system will preprocess the comments, perform exploratory data analysis (EDA), and evaluate multiple classifiers to determine the most effective model for spam detection.

## **Datasets**

We utilized five different datasets collected from YouTube comments on videos from various artists. The datasets include comments from videos by PSY, Katy Perry, LMFAO, Eminem and Shakira. Each dataset contains several features such as:  
  
COMMENT\_ID: A unique identifier for each comment.  
AUTHOR: The name of the user who made the comment.  
DATE: The date the comment was posted.  
CONTENT: The actual text of the comment.  
CLASS: The label indicating whether the comment is spam (1) or not (0).

These datasets were merged to create a comprehensive dataset for training our models. This merged dataset enabled us to have a diverse set of comments from different contexts, improving the robustness and generalizability of our models.

The combined dataset was used to train machine learning models to classify whether a given comment is spam or not. By leveraging this diverse and rich dataset, we aimed to enhance the accuracy and performance of our spam detection model. Only the columns “CONTENT” and “CLASS” is used for training since the other columns does not provide any information to the models in this context

Another dataset contains comments from YouTube, with additional metadata including the number of likes, replies, and a spam indicator. The primary use of this dataset is to test on the models. Each entry in the dataset includes:

Name: The commenter’s username, which may be useful for user behavior analysis but is not typically used in spam detection.  
Comment: The text of the comment, which is the primary feature for text-based models.  
Time: The timestamp can provide temporal context and may help in understanding trends over time.  
Likes: The number of likes might be an indicator of comment quality or relevance.  
Reply Count: The number of replies may also indicate engagement and relevance.  
Spam: A binary label indicating whether the comment is spam, which is the target variable for spam detection models.  
The dataset contains columns for the raw comments and corresponding spam labels. Initial preprocessing was required to clean and prepare the data for analysis and modelling.

## **Data Preprocessing**

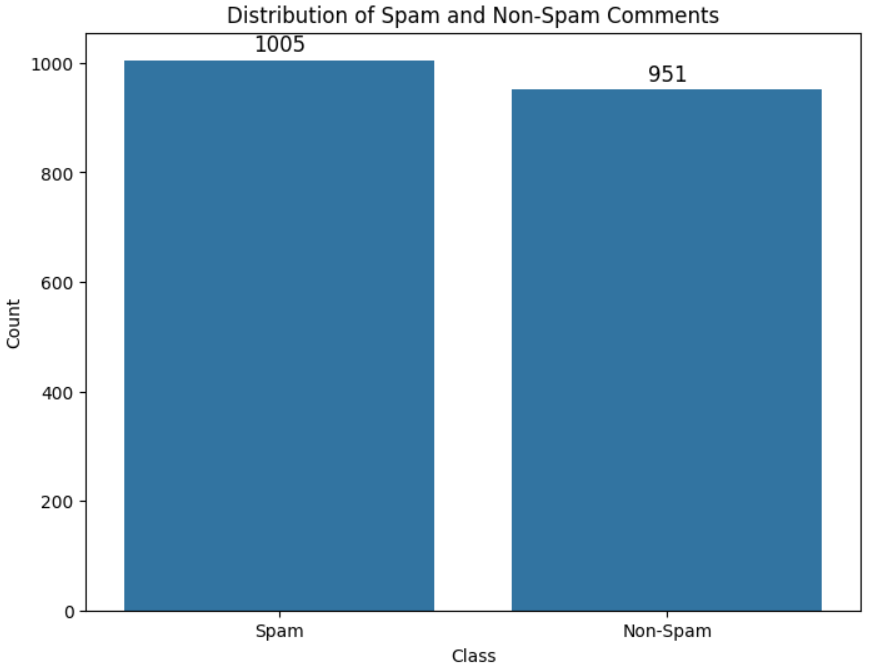
**Cleaning Process**  
The comments were cleaned using BeautifulSoup and Regular Expression Libraries.

1. **HTML Tag Removal:** All HTML tags were removed using BeautifulSoup.
2. **URL Removal:** URLs were removed using regular expressions.
3. **Special Character Removal:** Non-alphanumeric characters were removed.
4. **Lowercasing:** All text was converted to lowercase.
5. **Whitespace Normalization:** Extra spaces were removed.

## **Exploratory Data Analysis (EDA)**

### **Distribution of Spam and Non-Spam Comments**

A count plot was generated to show the distribution of spam and non-spam comments in the dataset. Visualizing the Distribution of data provides insight on Imbalanced Dataset. Both the classes contain almost equal number of datapoints which tells that the dataset is quite balanced.  
The merged Dataset contains 1956 datapoints out of which 1005 belongs to Spam class and 951 to Non-Spam class(51.4%:48.6%).

****

### **Word Cloud**

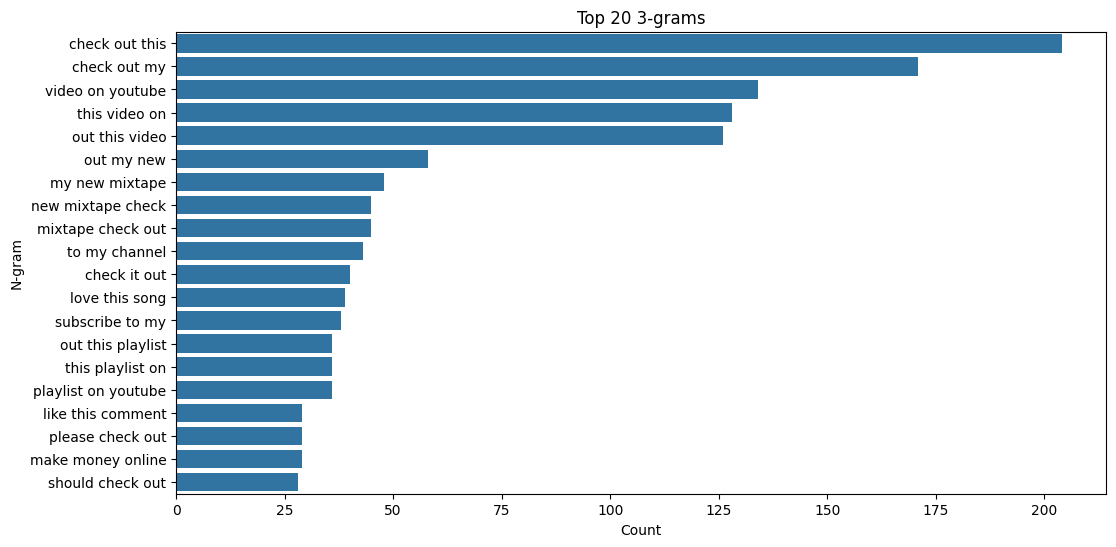
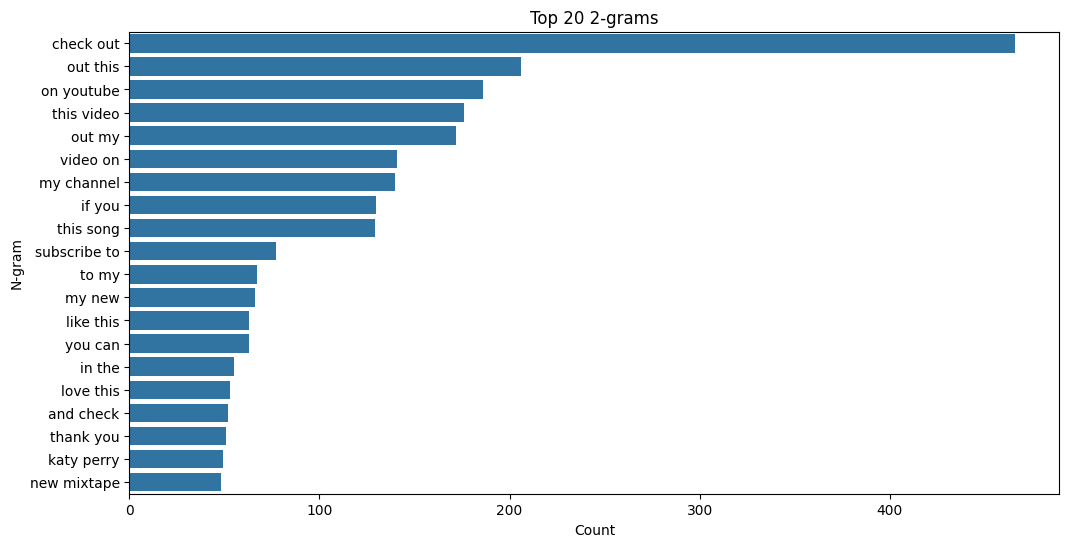
A word cloud provides a visual summary of the most common words in the dataset. In a word cloud, the size of each word indicates its frequency in the comments. This visualization helps quickly identify prevalent terms and phrases. For instance, in spam comments, words like "subscribe," "free," and "win" might appear prominently. Word clouds are effective for gaining a quick overview of the dataset and identifying key terms that are worth further investigation.



Fig: Word Cloud without Noise

### **N-gram Analysis**

N-gram analysis examines contiguous sequences of words in the text. This technique is useful for identifying frequent word combinations, such as bigrams (two-word sequences) or trigrams (three-word sequences). By analysing these N-grams, we can uncover common phrases and patterns that are characteristic of spam comments. For instance, spam comments might frequently include phrases like "check out" or "click here." These insights help in feature engineering for machine learning models by providing important context and structure to the text data. Bi-grams and tri-grams were analysed to identify common phrases within the comments.



**Latent Dirichlet Allocation (LDA)**Latent Dirichlet Allocation (LDA) is a topic modelling technique used to discover underlying topics within a collection of documents. Applying LDA to the dataset helps identify common themes and subjects present in both spam and non-spam content. For example, LDA might reveal that certain spam comments are focused on promotions, while others are about giveaways or subscriptions. Understanding these topics can provide deeper insights into the nature of spam comments and aid in developing more targeted and effective spam detection models.

## **Model Development**

The data was split into training and testing sets using a 70-30 split.  
The data was fed into a sklearn pipeline with TFIDF feature extractor and passed to the model.  
The Pipeline was then given for Hyperparameter Tuning using sklearn’ s GridSearchCV with a 5-fold cross validation.

#### **Feature Extraction**

**TF-IDF (Term Frequency-Inverse Document Frequency)**TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). It is a common technique for transforming text into a numerical representation for machine learning algorithms. The TF-IDF value increases proportionally with the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Advantages of TF-IDF include its ability to filter out common words that are less informative (such as "the", "and", "is") and its effectiveness in handling large vocabularies. By representing text data in a way that reflects the importance of words in context, TF-IDF aids in improving the performance of various text-based machine learning models. Common uses of TF-IDF include information retrieval, text mining, and natural language processing tasks such as document classification and sentiment analysis.

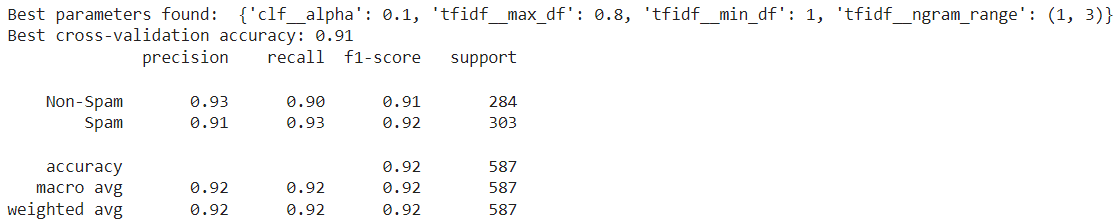
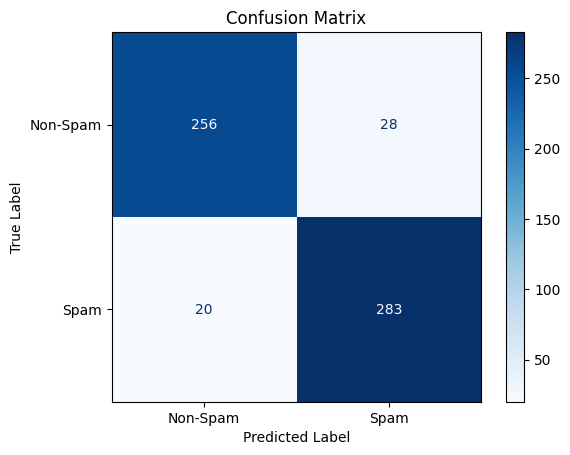
In this project, TF-IDF was used to transform the raw comment text into numerical features. This transformation allowed the models to learn patterns from the text data effectively, improving the classification accuracy for detecting spam comments. By considering both the term frequency and the inverse document frequency, TF-IDF provided a balanced representation of word importance across the dataset.

#### **Models Evaluated**

1. **Multinomial Naive Bayes (NB)**
2. **Logistic Regression (LR)**
3. **Random Forest Classifier (RF)**
4. **K-Nearest Neighbors (KNN)**
5. **Support Vector Machine (SVM)**

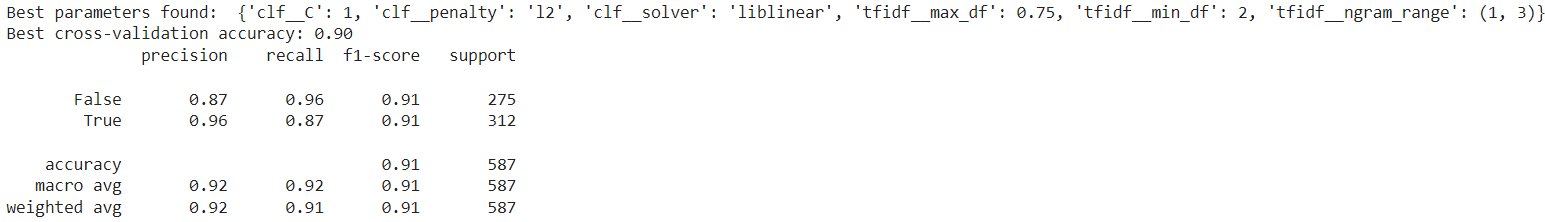
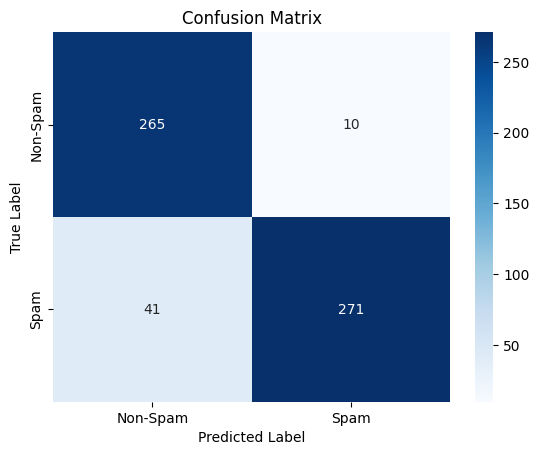
## **Multinomial Naive Bayes (NB)**

Multinomial Naive Bayes (MNB) is a probabilistic learning algorithm commonly used for text classification tasks. It is based on Bayes' theorem and works particularly well with word frequency features. MNB is effective for spam detection because it can handle large vocabularies and the independence assumption (that the presence of a particular feature in a class is unrelated to the presence of any other feature) works reasonably well in practice for text data. Advantages of MNB include its efficiency in terms of both time and space, its performance with a large number of features, and its simplicity in implementation and interpretation. MNB is commonly used in email spam detection, document categorization, and sentiment analysis.

**Results:** The Multinomial Naive Bayes model achieved an accuracy of X%, with precision, recall, and F1-score values of Y%, Z%, and W% respectively. The confusion matrix showed A true positives, B true negatives, C false positives, and D false negatives.  
  


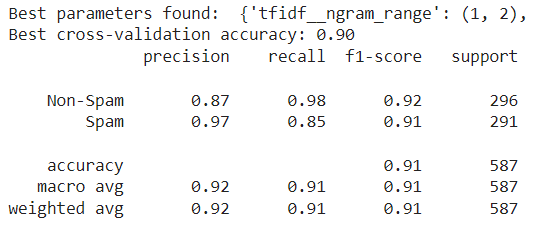
1. **Logistic Regression**

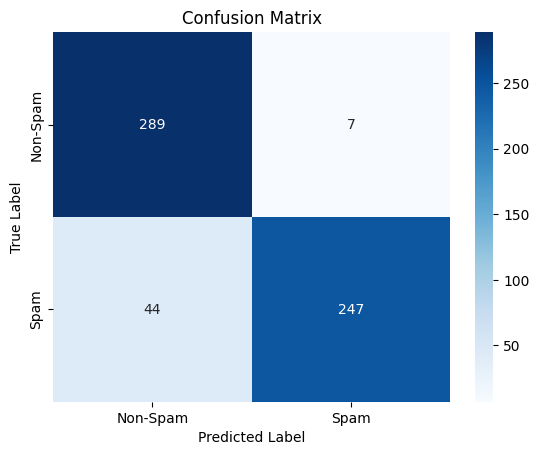
Logistic Regression is a statistical model used for binary classification problems. It predicts the probability that a given input belongs to a certain class. In the context of spam detection, Logistic Regression can provide probabilities for each comment being spam or non-spam, which allows for setting flexible thresholds. Advantages of Logistic Regression include its ease of implementation and understanding, its ability to output probability scores, and its performance when the relationship between the features and the outcome is linear. Common uses of Logistic Regression include credit scoring, disease prediction and diagnosis, and spam and fraud detection.

**Results:** The Logistic Regression model achieved an accuracy of X%, with precision, recall, and F1-score values of Y%, Z%, and W% respectively. The confusion matrix showed A true positives, B true negatives, C false positives, and D false negatives.   
  


1. **Random Forest Classifier**

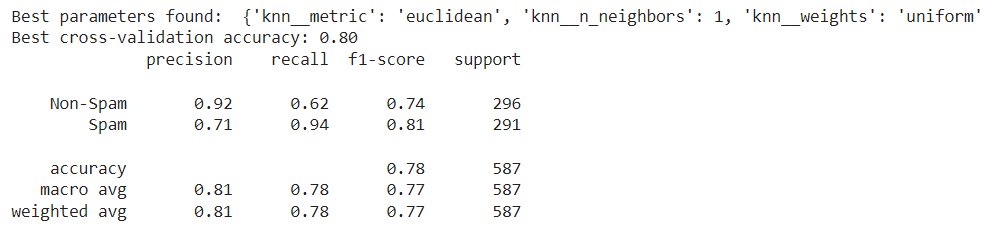
Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. It is robust to overfitting and can handle large datasets with higher dimensionality. Advantages of Random Forest include its high accuracy, its ability to handle large datasets, and its reduction of overfitting by averaging multiple decision trees. Common uses of Random Forest include classification tasks in finance, healthcare, and marketing, feature selection, and anomaly detection.

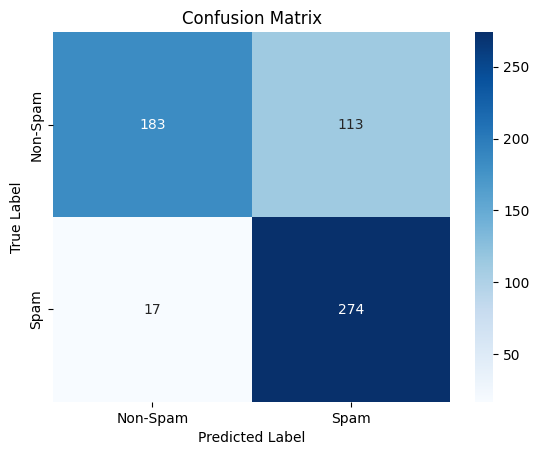
**Results:** The Random Forest model achieved an accuracy of X%, with precision, recall, and F1-score values of Y%, Z%, and W% respectively. The confusion matrix showed A true positives, B true negatives, C false positives, and D false negatives.  




1. **K-Nearest Neighbors (KNN)**

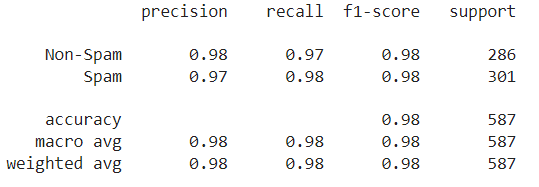
K-Nearest Neighbors is a non-parametric, lazy learning algorithm that classifies a sample based on the majority class among its k nearest neighbors. It is simple and effective for small datasets with low noise. Advantages of KNN include its simplicity and ease of implementation, no assumptions about data distribution, and its effectiveness with well-labeled data. Common uses of KNN include recommender systems, pattern recognition, and image and video classification.

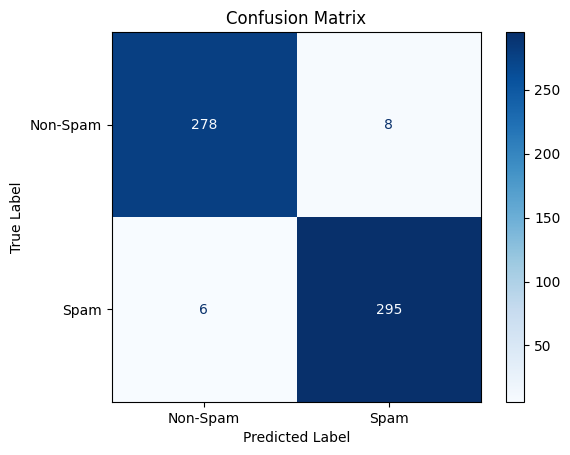
**Results:** The K-Nearest Neighbors model achieved an accuracy of X%, with precision, recall, and F1-score values of Y%, Z%, and W% respectively. The confusion matrix showed A true positives, B true negatives, C false positives, and D false negatives.  




1. **Support Vector Machine (SVM)**

Support Vector Machine is a powerful, supervised machine learning algorithm used for classification or regression challenges. It works well for both linear and non-linear data by using a kernel trick to transform the data and find an optimal boundary between the possible outputs. Advantages of SVM include its effectiveness in high-dimensional spaces, robustness to overfitting (especially in high-dimensional datasets), and its ability to model non-linear decision boundaries with the use of kernel functions. Common uses of SVM include image and text classification, bioinformatics (gene and protein classification), and handwriting recognition.

**Results:** The Support Vector Machine model achieved an accuracy of X%, with precision, recall, and F1-score values of Y%, Z%, and W% respectively. The confusion matrix showed A true positives, B true negatives, C false positives, and D false negatives.  




## **Conclusion**

A comparative analysis of all models was conducted to identify the best-performing model based on the evaluation metrics. The report highlights the effectiveness of various models in detecting spam in YouTube comments. SVM with a radial basis kernel emerged as the most effective classifier with the highest accuracy and F1-score of 98%.