**Project Name:** Youtube Comments Spam Detection

**Github Link:** https://github.com/projectsforstudents2022/Youtube\_Comments\_Spam\_Detection.git

**Why was this project created?**

As online social networks have improved in quality, spammers have discovered that they can easily use them to trick users into engaging in harmful behavior by leaving spam comments on videos. In this effort, spam detection is done and YouTube comments are taken.

**What problem is it solving?**

The goal of the project is to simply describe the machine learning techniques and the prediction technique. Using the proposed categorization method, the spam comment is anticipated. The negative viewpoint of the information contained in the submitted videos will be ruined by the offensive spam comments.

**Entire explanation of project**

* **PROPOSED APPROACH**

Importing a dataset and Preprocessing of the communications is required before preparation may begin. First, lowercase letters must be used for all characters. The word that appears in both capital and lowercase must be treated as one and the same word, not as two distinct words. Then, each message in the data collection must be tokenized. The primary benefit of incorporating the terms in the dataset is that it can decrease the amount of uncertainty in the final results forecast since those phrases have a notable impact on frequency counts in spam and ham comments on YouTube. A supervised characteristic called attribute significance ranks characteristics according to how well they predict an objective.

After preprocessing, there must be a method of creating a version that maintains the project's functional capabilities in line with the labelled model, which is created in accordance with the Supervised set of rules. For the purpose of predicting binomial or multinomial values of a variable, logistic regression is used. To determine the result, a statistical method is employed. The result is of a binary nature. Since it follows the Bernoulli distribution and employs the logit function to predict the likelihood of a binary event, the outcome will be either x or y. Here, it analyses the dataset and forecasts whether x or y is spam or ham.

Algorithm for creating next word prediction model :

**Step 1:** Dataset is imported

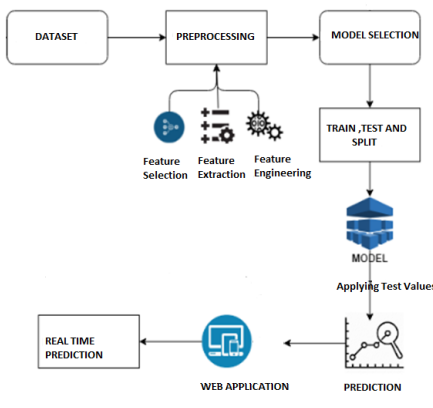
**Step 2:** Data Cleaning

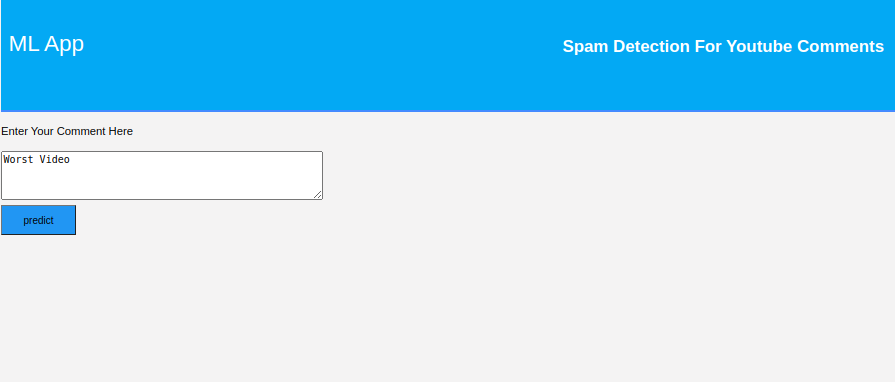
**Step 3:** Label Encoding

**Step 4:** Model Building (Logistic Regression)

**Step 5:** Train Model

**Step 6:** Prediction

* **DATA FLOW DIAGRAM**
* **RESULT**



* **CONCLUSION**

There are many methods used to categorize YouTube comments as spam and non spam. This method has been put to the test using real-time YouTube comments, and the results show that it is 84% accurate overall.