**TEXT CLASSIFICATION**

**“SPAM EMAIL CLASSIFIER”**

**1. Problem Definition**

* **Goal:** Classify emails as either **spam** or **not spam** (ham).
* **Importance:** We can be saved in trapped done through emails if we already know that the email is spam or not. If it is spam so it means it is fraud ,so we ignore that email.

**2. Data Collection**

* I collect the dataset through Kaggle Dataset .

**3. Data Pre-Processing**

1. **Text Cleaning**

* Remove extra spaces
* Convert text into lower case
* Remove HTML tags
* Remove stopword like is,of,the and word which are not important

1. **Tokenization**

* Convert sentences into different different tokens(words).

1. **Stemming/Lemmatization**

* Convert the word into their base root form (walking--walk)

**4 . Feature Extraction**

* We did feature extraction to convert the word into vectors so that the model can understand data easily and perfoming well.

1. **Bag of words**

* They firstly check the frequency of each words then they started vector representation.

1. **TF-IDF**

* It gives the the first priority to the word which are rarely used and give the less priority to the word which are continuous used.

**5 . Splitting the Data**

* In these step we splitting our dataset into training and testing and then run the model on training data and then check for testing data and find which model are more accurate.

**6 . Model Selection**

1. Logistic Regression
2. Naïve Bayes
3. Support Vector Machine(SVM)