**Problem Statement**

To build a classification model using Natural Language Processing techniques to identify if an email is a spam based on the transcripts of the email.

**Architecture**



**Steps performed on data**

**Step 1:** Downloaded the Spam classifier data set from Kaggle from the link below

<https://www.kaggle.com/uciml/sms-spam-collection-dataset>

**Step 2:** Imported the dataset into pandas dataframe and validated/cleaned the data and keep the relevant features. The only features required in this project are the messages and labels. The messages are the transcript of the columns and labels identify if the message is spam or HAM

**Step 3:** Applied Lemmatization technique on the Message column of the dataset and removed stop words and special characters. NLTK library’s WordNetLemmatizer() was used and output is stored in a list

*Note: Even stemming can be applied here, but I chose Lemmatization.*

**Step 4:** Bag of Words and TFIDF algorithm is applied to the Lemmatized data and the predictor data set X\_BOW and X\_TFIDF is created storing results of Bag of Words and TFIDF algorithm respectively.

**Step 5:**  Hyperparametertuning is applied to each dataset using the follow ML classification algorithms

1. Logistic Regression
2. Naïve Bayes Classifier
3. Naïve Bayes Gaussian
4. Random Forest Classifier
5. Decision Tree Classifier
6. SVM
7. XG Boost

**Step 6:** Best estimator and best dataset for the model (TFIDF or BOW) is chosen from Hyperparametertuning.

**Step 7:** The best dataset is taken, and train test split is done, and the Best estimator fit is applied to it.

**Files to run the model**

1. Spam Classifier.ipynb
2. Spam Classifier.py
3. Spam.csv