**ML ALGORITHMS IN PYTHON FOR HAM AND SPAM MESSAGE CLASSIFICATION**

**PROBLEM STATEMENT:**

Email is one of the most commonly used modes of communication in this modern era for Education, Banking, and Advertisement etc. As the technology is advancing cyber crime has also increased. Spam mail means unwanted mail or unused mail, that means there are no use in present and not in future. Ham is opposite of spam means wanted and useful mail. There are mails containing advertisement of some commercial websites for purchasing their products. As they are from unknown sources our inbox get filled with very huge amount of spam messages. So to overcome with we will prepare a model that will categorize the messages received by our devices as spam or ham. In order to achieve this, data from the messages is to be collected first and natural language processing techniques are to be applied on it. This spam filtering technique will help the mobile user to have better visualization of the inbox. Unnecessary mails will be marked as spam so mobile user need not to waste their time reading them.

Classifying spam and ham messages is one of the most common natural language processing tasks for emails and chat engines. With the advancements in machine learning and natural language processing techniques, it is now possible to separate spam messages from ham messages with a high degree of accuracy.

**DATA FLOW DIAGRAM:**

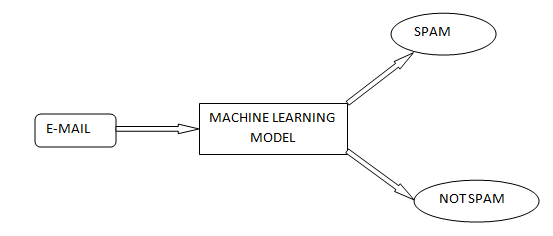
****

Figure: EMAIL SPAM CLASSIFIER USING ML MODELS

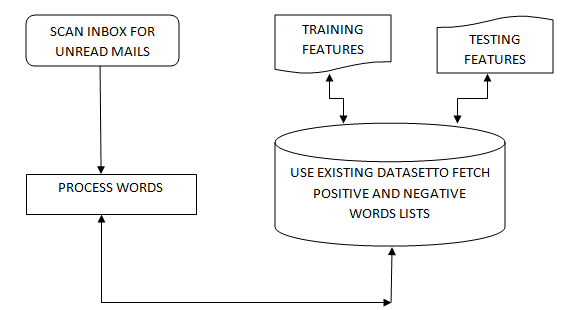
****

Figure: WORD PROCESSING AND CLASSIFICATION FOR TRAINING USING EXISTING DATASET

**DATASET:**

We will treat ham and spam message classification as a supervised machine learning problem. In a supervised machine learning problem, the inputs and the corresponding outputs are available during the algorithm training phase. During the training phase, the machine learning algorithm statistically learns to find the relationship between input texts and output labels. While testing, inputs are fed to the trained machine learning algorithm which then predicts the expected outputs without knowing the actual outputs.

For supervised ham and spam message classification, we need a dataset that contains both ham and spam messages along with labels that specify whether a message is a ham or spam.

All the Machine Learning Algorithms works on two stages:-

* Training Stage.
* Testing Stage.

So In the Training Stage Naive Bayes creates a Lookup table in which they store all the possibility of probability which we are going to use in the Algorithm for predicting the result.

And In the testing phase let suppose you have given a test point to the algorithm to predict the result, they fetch the values from the lookup table in which they store all the possibility of probability and use that value to predict the result.