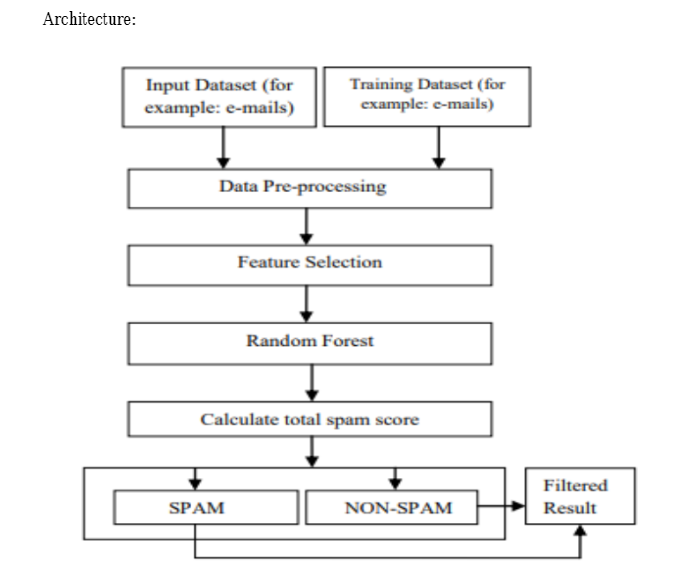
I would like to take the help of random forest algorithm , for classifying unlabelled data in the form of paragraphs , messaging and emails into spam and non - spam. For this , the following architecture should be followed :   


Pre-processing of data :

Pre-processing is a critical stage in text mining. Tokenization, stop word removal, and stemming are the three phases of preprocessing activities for message categorization. Tokenization is the initial stage. All symbols (@, #, %, and $), punctuation, and numbers will be deleted throughout the tokenizing process. The remaining strings will be tokenized. The second stage is to eliminate stopwords. Numerous commonly used English terms are worthless in information retrieval (IR) and text mining. These are referred to as 'Stop words.' Stop-words, which are functional terms peculiar to a language, are common words that provide no information (i.e., pronouns, prepositions, conjunctions). The common terms, which are the most often occurring words in a text, such as "we," "are," "is," and so on, are eliminated at this stage. There are around 400-500 stop words in the English language. The stop word list is based on the frequency of the words. By comparing the two lists, this method will determine which words fit the stop word lists. Eliminating these phrases frees up space for document storage and reduces the time required for the search process. The third stage is stemming. "Stemming" refers to the process of determining the origins of words and eliminating prefixes and postfixes. Stemming converts word forms such as adjectives, nouns, and verbs to

homologically similar words. For example, both "capturing" and "captured" are shortened to "capture."

After that , the feature selection takes place ,Feature selection is the process of evaluating data (such as a collection of typical emails) and identifying which features (words) would aid in classification the most. These features (words) are then used to train a classifier. TF is

one of the approaches to selecting features.

Then the Text Tokenization happens

We can use TF-IDF vectorizer (Term Frequency — Inverse Document Frequency) which basically takes the term frequency which is the frequency with which each word appears in a document and then makes use of that word to make sense of the relevance of the word to the document

Then we split the data an use train\_test\_split from sklear.model\_selection  
Random forest is a meta-learner made up of numerous individual trees. Each tree casts a vote on the overall classification of the provided data set, and the random forest method selects the classification with the most votes.

Each decision tree is constructed using replacement sampling from a random subset of the training dataset. That is, certain entities will appear multiple times in the sample, while others will not. Each decision tree is constructed using a model based on a separate random subset of the training dataset and a random subset of the available variables to determine how to divide the dataset optimally at each node. Each decision tree is constructed to its full

potential size without any pruning. When combined, the Random forest's decision tree models create the final ensemble model, in which each decision tree votes for the outcome and the majority wins.

The next step is to select a classifier. This phase will often entail the selection of numerous candidate classifiers and their evaluation against the testing set to see which performs the best.  
  
Evaluation Measures:

Spam classification evaluation is distinct from the evaluation of many other classification problems. Numerous prior

publications make extensive use of performance metrics such as precision, recall, Fscore, and accuracy. Precision is

defined as the percentage of accurate spam predictions. When precision is increased, the likelihood of incorrectly

identifying a legitimate message as spam decreases. Spam Recall looks at the percentage of spam emails that are

recognised. A larger spam recall highlights the decreased likelihood of misclassifying spam as authentic mail.

Simply said, the F-score is the harmonic mean of precision and recall. On the other hand, accuracy refers to the

proportion of spams and hams that are accurately classified.

from sklearn.metrics import accuracy\_score

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import classification\_report