b) Text Categorization

Categorization systems input a large flow of data like official documents, military casualty reports, market data, newswires etc. and assign them to predefined categories or indices. For example, The Carnegie Group's Construe system (Hayes, 1991) [54], inputs Reuters articles and saves much time by doing the work that is to be done by staff or human indexers. Some companies have been using categorization systems to categorize trouble tickets or complaint requests and routing to the appropriate desks. Another application of text categorization is email spam filters. Spam filters are becoming important as the first line of defence against the unwanted emails. A false negative and false positive issue of spam filters is at the heart of NLP technology, it has brought down the challenge of extracting meaning from strings of text. A filtering solution that is applied to an email system uses a set of protocols to determine which of the incoming messages are spam; and which are not. There are several types of spam filters available. *Content filters*: Review the content within the message to determine whether it is spam or not. *Header filters*: Review the email header looking for fake information. *General Blacklist filters*: Stop all emails from blacklisted recipients. *Rules Based Filters*: It uses user-defined criteria. Such as stopping mails from a specific person or stopping mail including a specific word. *Permission Filters*: Require anyone sending a message to be pre-approved by the recipient. *Challenge Response Filters*: Requires anyone sending a message to enter a code to gain permission to send email.

c) Spam Filtering

It works using text categorization and in recent times, various machine learning techniques have been applied to text categorization or Anti-Spam Filtering like Rule Learning (Cohen 1996) [27], Naïve Bayes (Sahami et al., 1998; Androutsopoulos et al., 2000; Rennie.,2000) [5, 109, 115],Memory based Learning (Sakkiset al.,2000b) [117], Support vector machines (Druker et al., 1999) [36], Decision Trees (Carreras and Marquez, 2001) [19], Maximum Entropy Model (Berger et al. 1996) [14], Hash Forest and a rule encoding method (T. Xia, 2020) [153], sometimes combining different learners (Sakkis et al., 2001) [116]. Using these approaches is better as classifier is learned from training data rather than making by hand. The naïve bayes is preferred because of its performance despite its simplicity (Lewis, 1998) [67] In Text Categorization two types of models have been used (McCallum and Nigam, 1998) [77]. Both modules assume that a fixed vocabulary is present. But in first model a document is generated by first choosing a subset of vocabulary and then using the selected words any number of times, at least once irrespective of order. This is called Multi-variate Bernoulli model. It takes the information of which words are used in a document irrespective of number of words and order. In second model, a document is generated by choosing a set of word occurrences and arranging them in any order. This model is called multi-nomial model, in addition to the Multi-variate Bernoulli model, it also captures information on how many times a word is used in a document. Most text categorization approaches to anti-spam Email filtering have used multi variate Bernoulli model (Androutsopoulos et al., 2000) [5] [15].

d) Information Extraction

Information extraction is concerned with identifying phrases of interest of textual data. For many applications, extracting entities such as names, places, events, dates, times, and prices is