This chapter describes prior and some important research work done to solve the problem of Spam E-mails. This chapter briefly reviews the important contribution done by the researchers in Origin based Filters, including Black-listing and White-listing of domain names / IP addresses and Challenges Response Systems. Also it reviews the Content based Filters which includes Machine Learning based Classifier and Semantic Similarity with Edge based Classifier. This review found to be very useful in comparing the results.

[Amlan Mohanty, 2011] have described the various amendments made to the IT (Amendment) Act, 2000. He has analyzed the legislative response to cyber crime in India with analysis of the Information Technology (Amendment) Act, 2008. The amendments of new crimes are examined by the author.

[Kigerl, 2014] have addressed the spammer behaviour, spam volume, spam compliance, spam locality with the CAN SPAM Act by analysing the 5,490,905 samples of spam emails received in the United States from March, 1998 to November, 2013.

[Govt. of India, 2013] have defined Spam as “the use of E-mail systems to send unsolicited bulk E-mails, especially advertising, indiscriminately. This notification has advice the government servants not to use their official E-mail addresses to subscribe on any unsafe or fake website.

[Ramasubramanian and Prakash, 2013] have presented a brief report which has discussed the historical growth, spread of spam and Internet abuse which includes telemarketing and mobile spam messaging, in India. They have also addressed the current and proposed Indian law which includes cybercrime.

[FTC, 2005] have presented some schemes which include the impact Spam, spyware which try to steals the user’s data and crashes their computers. Also, telemarketing and health claims which hits the weaknesses of users. The FTC in its report has recommended for making proper provisions to fight against these issues.
Chapter-2: Literature Survey

[IT Act, 2000] The Indian Government has passed an information Technology law to addressed the issues related to cyber crime., Information technology This has not addresss the problem of Spam.

[Amend_IT ACT, 2008] The amendments made to IT ACT, 2000 includes the issues rated to digital secure electronic signature, protecting data, provision of compensation in case of failure of data are made on 5th February, 2009.

[Coello, 2005] has studied the current technical and legislative solutions which are proposed by the governments and private companies to fight the problem of Spam.

[Robinson et.al, 2011] have presented a comparative study on legislative and non legislative measures used to combat identity theft and identity-related crime. This comparative study includes the definition and context of identity theft. They have summarised the overall legislation made by different countries.

[Karen Ng, 2005] have presented a review on various options for controlling the Spam in context with Canada. He has recommended two important points which should be considered in future Canadian Spam legislations, which government of Canada is competent to control and regulate the problem of Spam and have suggested to cover charter scrutiny under this law.

[Moustakas et.al, 2005] have presented an overview of different anti-spam laws including weaknesses of these laws, to fight against problem of Spam and have compared this with US and European Union anti-Spam Legislation.

[Hohlfeld et.al, 2012] have presented a work which find out the origins of the spamming process which concentrated on address harvesting on the web. They have suggested that, simple methods of obfuscations are still efficient for protecting addresses from being harvested.

[Ramachandran, 2006] have presented an evaluated the responsiveness of blacklisted IP addresses maintained.

[Erickson et.al, 2008] have carried out the experiments on the White list, they found that, it is really hard task for user to manage the white-list addresses. They have used C-R system to check the authenticated Seders.

[Martin et.al, 2005] have empirically analysed the behavioural features of Spam E-mail. These features can detect the abnormal activity of an E-mail and worm propagation.

CAPTCHA stands for ‘Completely Automated Public Turing test to Tell Computers and Humans Apart’. The concept of a CAPTCHA is motivated by real-world
problems faced by internet companies such as Yahoo and AltaVista. These companies offer free E-mail accounts, intended for use by humans. However, they found that many online vendors were using ‘bots’, computer programs that would sign up for thousands of E-mail accounts, from which they could send out masses of junk E-mail. Bots are misusing the free E-mail accounts offered by many companies.

[Haddadi and Zincir-Heywood, 2014] have evaluated eight different botnet traffic data sets. The results show that, the use of a flow exporter and a protocol filter indeed has an effect on the performance of botnet traffic classification. They have carried out experiments whose results show that the best performance is achieved using Tranalyzer flow exporter and HTTP filter using C4.5 classifier.

[Subramaniam et.al, 2010] have carried out the experimental analysis of machine learning techniques which are used for Spam E-mail filtering using content analysis which includes the Naive Bayesian classifier, Support vector machine and Neural Network. They have also carried out the survey of Anti-Spam legislation all over the world.

[Liu and Ting, 2012] have proposed online active multi field learning approach for Spam filtering. They have proved that, this proposed method outperforms text classification algorithms.

[Wang et.al, 2006] have carried out work to find performance of linear classifiers such as the Perceptron and Winnow on PU1 and LingSpam dataset. They have demonstrated that, these classifiers outperforms than Naive Bayesian Classifier.

[Zi-Qiang Wang, et.al, 2006] have addressed the problem of Spam filtering and have proposed a method for filtering Spam using support vector machine classifier after applying the process of feature selection. They have shown that, this method gives optimum results than the existing filtering techniques.

[Fatiha Barigou et.al, 2014] have proposed an modified K-Nearest Neighbours (KNN) method for Spam filtering which is labelled as Cellular Automaton Combined with KNN (CA-KNN). They have proved that, the proposed method outperforms the traditional k-NN classifier.

The Naive Bayesian is proposed by [Sahami et.al, 1998]. It is a statistical approach based on probability which is used to classify Spam or Ham emails.

[Pantel and Lin, 1998] have presented simple Spam filtering program ‘SpamCop’ which is able to identify 92% of spam while misclassifying 1.16% of non-spam E-
mails. ‘SpamCop’ treats an E-mail as a multi set of words and employed a Naive Bayes algorithm.

[Schneider, 2003] described algorithms with Naïve Bayes text classifiers in the context of anti-spam E-mail filtering using two statistical event model. They have introduced family of feature ranking function for feature selection in the multinomial event model.

[Garcia et.al, 2004] have investigated the effectiveness of several spam filtering techniques and technologies. They have analyzed the performance by simulating E-mail traffic under different conditions. They showed that, genetic algorithm based spam filters perform best at server level while Naive Bayesian filters are most appropriate for filtering at user level.

[Metsis et.al, 2006] have discussed five different versions of Naive Bayes and compared them on six new non-encoded datasets that, contain ham messages of particular Enron users and fresh spam messages.

[Graham, 2003] A Plan for Spam suggested a Bayesian filter is first trained on a set of known ham and a set of known spam. The text of each email is broken into tokens or tokenized. A token corresponds to a word or HTML tag. The number of times the token has been seen in spam and ham is stored. When the trained filter receives an email, it tokenizes the message and assigns each token a probability based on the frequency data. This is known as a word value. A Bayesian calculation then combines these probabilities into a final score which is used to classify the email as ham or spam.

[Hovold, 2005] have proved that, it is possible to achieve very good classification performance using a word-position based variant of Naive Bayes.

[Awad and ELseuofi, 2011] have reviewed some of the most popular machine learning methods including Bayesian classification, k-NN, ANNs, SVMs, Artificial immune system and Rough sets to the problem of spam Email classification using SpamAssassin corpus.

[Haskins and Nielsen, 2004] have proposed solution SpamAssassin to block Spam Emails. SpamAssassin includes Header analysis, Body analysis, Distributed checksums and Bayesian analysis.

[Paganini, 2003] have developed Active Spam Killer (ASK), a program that attempts to validate unknown senders before allowing delivery of their message. When an E-mail is received from an unknown user. ASK calculates the MD5 checksum of the
message plus a secret MD5 key which is configured during installation time. This number is sent as part of a ‘confirmation message’ back to the user. If the user replies to it, the confirmation number mentioned in the subject is recognized then message is dequeued and delivered. If the user does not reply, the message remains queued until it is removed.

[Yanhui Guo et.al 2003] have implemented a feedback mechanism for text classification systems which reduces the need for labelled training documents by unifying the strengths of k-NN and linear classifiers. It provides the advantage of continuous learning in the batch-adaptive text considerations, on the batch-adaptive filtering task.

[Wang et.al, 2006] have investigated the use of Hill Climbing, Simulated Annealing, and Threshold Accepting optimization techniques as feature selection algorithms. They have compared the performance of these techniques with the Linear Discriminate Analysis. Their results show that all these techniques can be used not only to reduce the dimensions of the E-mail, but also to improve the performance of the classification filter. Among all the strategies, simulated annealing has the best performance which reaches a classification accuracy of 95.5%.

[Lai, 2007] has carried out systematic experiments of machine learning techniques such as, SVM, k-NN and Naive Bayesian on different parts of an E-mail. His experiments show that, using the header part only one can achieve satisfactory performance and the idea of integrating disparate methods is a promising way to fight spam. He has carried out a comparative study on performance of these machine learning methods in spam filtering.

[Faure et.al, 2007] have devised new method to detect spam email. They have implemented a k-nearest neighbour’s algorithm to test their method.

[El-Halees, 2009] have compared six supervised machine learning classifiers which are maximum entropy, decision trees, artificial neural nets, Naive Bayes, support system machines and k-nearest neighbour. He has applied these methods on stemmed Arabic spam words. He showed that, for most cases, classifiers using feature selection techniques can achieve better performance than the filters which do not use it.

[Romero et.al, 2010] used blog comment corpus as a case study in which they have used 50 pages and 1024 blog comments. The percentage of spam of this corpus is 67%. They have applied and compared the result of four machine learning techniques
such as Naïve Bayes, K-nearest neighbour, neural networks and the support vector to classify blog comments as spam and non-spam.  

[Khan et.al, 2010] have reviewed important techniques and methodologies that are employed in text documents classification based on existing literature. They have focused mainly on text representation and machine learning techniques.  

[Rahmi, 2011] have extracted features with hybrid features selections of a publicly available test corpus. They have achieved 96% accuracy.  

[Caruana and Li, 2012] have reviewed emerging approaches to spam filtering built on recent developments in computing technologies which includes peer-to-peer computing, grid computing, semantic Web, and social networks. They have addressed a number of perspectives related to personalization and privacy in spam filtering. Finally they have concluded that, important advancements have been made in spam filtering in recent years, high performance approaches remain to be explored due to the large scale of the problem.  

[Askì, 2013] have proposed an algorithm to classify E-mails and minimize spam using nearest neighbor classifier. This approach involves low computational load in relatively high rate relying on a Hash table as well as a flag varying in the range of {1, 0}.  

[Uysal, 2013] have investigated the impact of several feature extraction and feature selection approaches on filtering of short message service (SMS) spam messages in two different languages Turkish and English. The entire feature set of filtering framework consists of the features originated from the bag-of-words (BoW) model along with the ensemble of structural features (SF) specific to spam problem. The distinctive BoW features are identified using information theoretic feature selection methods. Various combinations of the BoW and SF are then fed into widely used pattern classification algorithms to classify SMS messages. They have evaluated filtering framework on both Turkish and English SMS message datasets which is Turkish SMS message collection. They have demonstrated that, the combinations of BoW and SFs, rather than BoW features alone, provide better classification performance on both Turkish and English message datasets.  

[Pawlak, 1982] have presented an alternative to fuzzy set theory and tolerance theory using approximate operation on sets, equality of sets and on inclusion of sets  

[Chunping, 2013] have analyzed spam filtering technology by carrying out detailed study of Naïve Bayes algorithm, and proposed the improved Naïve Bayesian mail
filtering technology. He demonstrated that, improvement can be seen in text selection as well as feature extraction. He proposed a new feature selection method based on class conditional distribution algorithm and proved that, the proposed algorithm can effectively filter spam.

[Saab et.al, 2014] have presented a survey of some popular filtering algorithms that rely on text classification to decide whether an email is unsolicited or not. This algorithm is executed on the Spambase dataset to identify the best classification algorithm in terms of accuracy, computational time, and precision and recall rates.

[Peace et.al, 2015] have carried out the comparative analysis of prediction success using rapid miner tool. In this process of prediction, they have applied machine learning classifiers such as k-NN and ANN on English Premier League dataset. They have proved that, ANN outperforms k-NN with prediction success of 70%.

[Gomaa and Fahmy, 2013] have carried out survey of the existing techniques on text similarity by considering three major approaches such as String-based, Corpus-based and Knowledge-based similarities.

[Elavarasi et.al, 2014] have presented a comprehensive survey of semantic similarity measures with various approaches including path based measures, information based measures, feature based measures and hybrid measures. All these methods are discussed with their advantages, disadvantages, features and other issue.

[Kim et.al, 2007] have presented a new approach different from content based method which uses user preferences for constructing an anti-spam mail system. They have constructed user preference ontology using important concept. Using inference engine they have proved that their method gives good performance.

[Kiamarzpour et.al, 2013] have proposed new method for classifying the spam E-mail. They have showed high accuracy of E-mail classification by using the several decision trees in combination with ontology.

[Ted Pedersen, 2010] has presented an empirical analysis using similarity measures for the pairs of concepts using Information Content.

[Sánchez et.al, 2013] have presented a survey of the ontology-based approaches. They have presented new ontology based measure by exploitation of taxonomical features. They have proved that this method provides a high accuracy.

[Polcicova and Navrat, 2002] have used content based filtering for representing new items by using the word clusters based on their semantic similarity.
[Li et.al, 2006] have proposed an algorithm which uses semantic information and word order information present in the sentence. This method can be useful for text knowledge representation and discovery. He proved that, the proposed similarity measures has significant correlation to human intuition.

[Youn and McLeod, 2007] have proposed ontology based efficient spam E-mail filtering method which is adaptive in nature.

[Lin and Sandkuhl, 2008] have presented an overview of string similarity, synonyms, structure similarity and based on instances used in current ontology matching systems. They have proved that, the synonyms can help to solve the problem of using different terms in the ontologies for the same concept. Also, it proved that, WordNet thesauri can be used to improve the similarity measures.

[Meng et.al, 2013] have devise an algorithm for semantic similarity metric of word pairs which has given effective performance, which is using path length as well as IC values of two words into consideration. Their experimental analysis has shown that, the coefficient of the proposed algorithm with human judgment is 0.8820.

[Lingling Meng et.al, 2013] have carry out the survey of semantic similarity measures which includes path based measures, information based measures, feature based measures and hybrid measures. All these are compared using features and performance. They have also discussed the advantages and disadvantages of these measures.

[Philip Resnik, 1995] have proposed a novel technique to measure the semantic similarity based on Information Content (IC) in ‘is-a’ taxonomy. He proved that, this new measure outperforms the traditional edge counting approach.

[Philip Resnik, 1999] have presented an algorithm with its effective analysis which has an advantage of taxonomic similarity for the purpose of resolving semantic as well as syntactic ambiguity. This method find the semantic similarity with shared Information Content in ‘is-a’ taxonomy.

[Youn and McLeod, 2007] have proposed an efficient spam E-mail filtering method which is adaptive in nature. They have crate an ontology using RDF data model. This data model is then used for checking the Spam E-mail using adaptive ontology

[Jiang and Conrath, 1997] have presented new approach for measuring semantic similarity or distance between words and concepts.
[Wu and Palmer, 1994] have concentrated on the semantic representation of verbs with the impact of problem of lexical selection. They have proposed a novel technique which is useful in correct lexical selections by searching the inexact matches. [Rodriguez and Egenhofer, 2003] have presented a method in which similarity function using which similar entity can be find out based on set of synonyms, semantic neighbourhood, and some features which includes parts, functions, and attributes. [Euripides et.al, 2006] have provided the result of empirical analysis of approaches which are used to calculate the semantic similarity of medical terms and natural language terms using MeSH and Wordnet ontology respectively.