Chapter 1. Introduction

Nowadays, email is an essential and useful tool for enabling rapid and cheap communication. It is a popular medium and an important part of the life (Whittaker, Bellotti, & Moody, 2005). However, spam or unsolicited email is a constant nuisance in the email communication. It can be an advertisement or some such explicit content that may have malicious code embedded in it. A study has estimated that 70 percent of all business emails are spams (Aladdin Knowledge Systems). The rapid growth of spam causes some serious hitches such as over-flowing users’ mailboxes, compromising some important emails, consuming bandwidth and storage space, as well as consuming user’s time to sort the emails (Lai, 2007).

At present, there is enhanced interest of the researcher in spam classification due to the complexity introduced by the spammers which makes difficult to distinguish between a spam (unsolicited emails) and ham (legitimate emails). Complexity may be due to attacks like Tokenisation (splitting or modifying a feature such as ‘free’ written as f r 3 3) and Obfuscation (hiding features by adding HTML or some other codes such as ‘free’ coded as fr&#101xe or FR3E), that alter the information on particular features (Goodman, Cormack, & Heckerman, 2007).

A number of email spam filtering methods are used for spam classification. A spam filter identifies spam email and prevents its delivery to the mailbox. Filters are used to mitigate the adverse impact of spam email and can operate like a reliable and predictable tool to eliminate unwanted emails though with a small risk of misclassification or removal of legitimate emails. In the research, many spam filtering methods have been tested but none of them was found perfect. However, filters are useful for an email recipient who has to endure the burden of identifying spam. The cost incurred to read every email for detecting spam is more than just consumption of time (Caliendo, Clement, Papies, & Scheel-Kopeinig, 2008). The
necessity of spam filter is increasing because it is difficult for a user to identify legitimate email by just reading the subject or the content. Sometimes, filters also make mistakes but they are used in conjunction with users to minimize these errors.

1.1 **Spam Emails**

Spam is not just unwanted email. It is a message that may include a payload, delivered by a sender despite the wishes of the recipient. The payload could vary from a number of schemes being advertised, bait for fraud or content which could have some malicious code in between. Such senders are commonly known as spammers (Gyongyi, & Garcia-Molina, 2005). Spammer gains some benefits from activities like money making, control on computers or access to personal information of a user. According to the definition proposed by Spam Track at the Text Retrieval Conference, a Spam is -

- “*Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient.*” (Cormack, & Lynam, 2005)

No precise definition of Spam Email has been found in the literature. Most of the spam emails are termed as “Unwanted Email” that is not suitable for all spam emails. Another suggested name for spam is “Unsolicited Commercial Emails” (Zdziarski, 2006), but this cannot be completely true because spam is not just an advertising material. Spam can also be represented as “Junk Email”, but it raises another question “What is a Junk Email?” Although most of the users are familiar with word “spam” and “spamming”, but a precise definition is yet to be produced.

Wikipedia, which is known as the biggest internet encyclopaedia, has given a definition of spam and spamming:
• “E-mail spam involves sending nearly identical messages to thousands (or millions) of recipients.” (Wikipedia)

And the definition for Spamming is:

• “Spamming is the abuse of any electronic communications medium to send unsolicited messages in bulk.” (Wikipedia)

Further Gordon V. Cormack (2007) has generalised the email spam definition given by Spam Track to capture the essential nature of spam as:

• “Spam is an unwanted communication intended to be delivered to an indiscriminate target, directly or indirectly, notwithstanding measures to prevent its delivery.” (Cormack, 2007).

The generalised definition of email spam can be applicable for different communication media such as text and voice messages (Cormack, Hidalgo, & Sánz, 2007; Dantu, & Kolan, 2005: Gómez Hidalgo, Bringas, Sánz, & García, 2006), social networks (Zinman, & Donath, 2007), or blog comments (Cormack, Gómez Hidalgo, & Sánz, 2007; Mishne, Carmel, & Lempel, 2005). It also applies to the web spam, which uses a search engine as its delivery mechanism (Castillo, Chellapilla, & Denoyer, 2008; Webb, Caverlee, & Pu, 2006).

![Fig 1. Different Kinds of Spam Sets (Gulyás, 2006)]
1.2 **Evolution of Spam Emails:**

The history of email spam can be categorised into three periods:

- Earlier, when spam letters were addressed and sent manually.
- Later, when various Machines were used for sending bulk emails.
- At last, after the development of Machine learning filters.

**1.2.1 The Early Years (Manual Spamming)**

This era of spamming started in 1978 and lasted till mid-90. At this time, the spams were sent manually. This process needed enamours amount of human effort so that it was difficult to sent spam to millions of user. In this phase, spammers sent spam one by one to the inner address lists of small communities.

**1.2.2 The Second Phase (Machines for Spamming)**

This phase, where a programmer was hired for making programmes for spamming, started in 1994. At this time, the amount of spam letters had uncontrollably increased. The first famous spammer in this phase was Jeff Slaton, known as “The Spam King”. This man was not only a good spammer but also charged $5 to the user for removing the email address from all spamming lists.

Slowly, spamming became a business. In 1995, the first known email address list with two million email addresses was released for sale. In 1997, the first real spam filtering software had been produced but it was unable to stop the exponential growth uncontrollable spam letters. Different tactics were used for tackling the tricks of spammers and a number of letter based filters were developed but unfortunately, spammers had the capability to change the spam word for beating such filters.
1.2.3 Third Phase (Machine against Machine)

A significant success against spam was achieved in 2002, when Paul Graham wrote and published an article “A Plan for Spam” (Graham, 2003). In this article, he discussed various machine learning methods and a statistical rule for classifying emails and introduced Bayesian network method. He wrote “I think it's possible to stop spam, and that content-based filters are the way to do it. The Achilles heel of the spammers is their message. They can circumvent any other barrier you set up. They have so far, at least. But they have to deliver their message, whatever it is. If we can write software that recognizes their messages, there is no way they can get around that”.

After the introduction of the Bayesian method, spam filtering has not only turned out to be more efficient but captured the semantic relationship between the words of an email for learning the email characteristics of users. This solution created new problems for spammers. Now, they were reluctant to test the filters before sending emails because these filters worked with a diverse knowledge base of the user.

1.3 Types of Spam

Spam is written in English language and can be divided at least into three categories.

- Advertisement Spam
- Financial Spam
- Phishing Spam

Advertisement Spam can also be subcategorized into different types. Most of the Spam follows similar patterns and could easily be identified. Over 99% Spam falls in the following categories:
1.3.1 Advertisement Spam

Majority of spam are used for commercial purpose where a product is offered directly to the consumer. It is a cheaper medium than any other advertising medium for promoting a product. Advertisement Spam can be further divided into following categories:

- **Marketing Spam:** It is difficult to identify the purpose of spam. A decade ago, the mechanism and purpose of sending a payload or spam were transparent. Most of the spams were sent by “Cottage Industries” to promote their products. Such spams were called “Marketing Spam”.

- **Online Pharmacy Spam:** Such spams are used to promote various versions of drugs like Viagra, Cialis, and Anti-depressant pills that can be purchased online.
- **Stock Encouraging Spam**: These spams are used to encourage people to buy cheap stocks.

![Email example](image)

*Fig 3. Online Pharmacy Spam*

*Fig 4. Stock Encouraging Spam*
• **Pornographic or (sex-) Dating Spam:** Sex or pornographic dating is often promoted by spam emails. Nowadays, the frequency of these spams has fortunately reduced in comparison to others.

![Fig 5. Pornographic or (Sex-) Dating Spam](image)

• **Pirate Software spam:** These spams are used to sell pirated software for a very cheap price compared to the original.

![Fig 6. Pirate Software Spam](image)
- **Online Casino spam**: Online casinos gambling are promoted through such spams.

![Email Example](image1.png)

Fig 7. Online Casino Spam

- **Fake Degrees spam**: Fake degree and diploma are sold by these kinds of email spams.

![Email Example](image2.png)

Fig 8. Fake Degree Spam
• **Mule Job Spam:** Such spams are used to promote jobs such as “Working from home”. These are a type of scam or mule jobs which are developed for money laundering.

![Figure 9. Mule Job Spam](image)

• **Cause Promotions Spam:** Such spams carry chain letter messages and used to promote some cause or event.

![Figure 10. Cause Promotions Spam](image)
1.3.2 Financial Spam

These spams are more dangerous than advertisement spams. In advertisement spam, at least a user can get something by paying money but in financial spam, spammers only attempt to fool the users and get their money without giving anything in return. Some types of financial spams are categorised below:

- **Fraud based Spam:** These spams have some request and emotional content which force the user to provide help for recovering millions of dollars from a foreign bank account, for example, “Nigerian Spam” that prompts one to send bank transaction information.

Fig 11. Fraud based Spam

- **Lottery Spam:** These spams include message like “You have already won some million $ etc.” which force a user to send some transaction fees for getting a lottery price.
• **Virus Spam:** Dangerous viruses are sent with attached executable files that can make the system of a user vulnerable. These messages are written to get all the personal information of a user or for spoofing purpose.
1.3.3 Phishing

The word “Phishing” evolved from the word “fishing” which indicates a fraudster fishing out personal information of a user. These spams are fake alert emails and appear to be sent from banks (mostly Citi Bank), PayPal, EBay etc. Such messages force the user to validate, confirm or monitor the details of the email to get personal information like bank account number. These spams are usually linked to a fake site which is used to make the users’ system vulnerable to fraud or steal money / goods and personal id.

They also make a user wary towards online shopping portals and hence cause a big loss for online businesses. Some methods of phishing are given below:

- **By the use of Company image:** For fishing the user, spammers use the logos, similar font and colour of famous companies.

- **Real company sites’ link:** Sometimes spammers use some real companies’ sites, together with fraudulent phishing sites, to lure the user to make mistakes.

![Phishing Spam](image)

Fig 14. Phishing Spam
1.3.4 Image Spam

Nowadays, to counter content-based spam filtering systems such as Machine Learning, Image spams are used. Content-based filters work with features/words of the email content but it’s hard to capture features embedded within an image file. Various sophisticated image recognition tools like Optical Content Recognition (OCR) and image processing are employed to capture the content of the image file.

![Image Spam](image)

Fig 15. Image Spam

1.4 Spam Consequences

Spamming is economically done because there is no operating cost for the senders and it’s hard to fix accountability for mass mailing. The number of spammers and the volume of the unwanted emails is enamours because of the low entry barrier. Estimation shows that the number of spam in 2011 was around seven trillion. Now spam has become a subject of many jurisdictions (Spamhaus Project). A number of consequences have been found from email spams (Caliendo, Clement, Papies, & Scheel-Kopeinig, 2008) which are discussed below:
• Spam generates an unregulated communication medium which is used to defraud targets without any permission, for selling sloppy products, installing computer virus etc.

• Network resource consumption is the other consequence. A large part of today’s email network traffic is spam that creates serious problems such as waste of bandwidth and storage, untimely / delayed delivery of emails and engulfing of some good emails etc.

Table 1. Statistics of Spam (Azam, 2002)

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily Spam emails sent</td>
<td>12.4 billion</td>
</tr>
<tr>
<td>Daily Spam received per person</td>
<td>6</td>
</tr>
<tr>
<td>Annual Spam received per person</td>
<td>2,200</td>
</tr>
<tr>
<td>Spam cost to all non-corporation Internet users</td>
<td>$255 million</td>
</tr>
<tr>
<td>Spam cost to all U.S. Corporations in 2002</td>
<td>$8.9 billion</td>
</tr>
<tr>
<td>Email address changes due to Spam</td>
<td>16%</td>
</tr>
<tr>
<td>Annual Spam in a 1,000 employee company</td>
<td>2.1 million</td>
</tr>
<tr>
<td>Users who reply to Spam email</td>
<td>28%</td>
</tr>
<tr>
<td>Users who purchased from Spam email</td>
<td>8%</td>
</tr>
<tr>
<td>Corporate email that is considered Spam</td>
<td>15-20%</td>
</tr>
</tbody>
</table>

Table 2. Categories of Spam (Azam, 2002)

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Products</td>
<td>25%</td>
</tr>
<tr>
<td>Financial</td>
<td>20%</td>
</tr>
<tr>
<td>Adult</td>
<td>19%</td>
</tr>
<tr>
<td>Scams</td>
<td>9%</td>
</tr>
<tr>
<td>Health</td>
<td>7%</td>
</tr>
<tr>
<td>Internet</td>
<td>7%</td>
</tr>
<tr>
<td>Leisure</td>
<td>6%</td>
</tr>
<tr>
<td>Spiritual</td>
<td>4%</td>
</tr>
<tr>
<td>Other</td>
<td>3%</td>
</tr>
</tbody>
</table>
• Spam also consumes human resources because it is an unpleasant experience and consumes a considerable amount of time to read, sort and delete from a mailbox. Some smart email spam may take extraordinarily long time because that appears as a legitimate email and difficult to filter with the summary information given by the user.

1.5 **Spam Filters**

A spam filter is a computer program that filters email spams. After detecting a spam, a filter labels it, moves it from the mailbox to a spam folder or simply deletes it. A number of spam filtering methods have been developed that can be statistically differentiated from each other. Gordon V. Cormack (2007) has given an appropriate definition for spam filter:

“Spam filter is an automated technique to identify spam for the purpose of preventing its delivery” (Cormack, 2007).

1.6 **Spam Filtering Methods**

Spamming can be done easily by the spammers. Various websites offer a number of necessary tools to the clients for sending “bulk emails”. In addition, they also provide the lists containing millions of valid email addresses for sending spam. A number of Non-technical and Technical solutions have been proposed for dealing with this. Some of the technical solutions have been implemented and are predicted to be good. On the other hand, Nontechnical solutions are difficult in implementation but have provided good ideas. The descriptions and various types of these solutions are discussed below.
### 1.6.1 Non-Technical Solutions

Non-Technical solutions are prepared by identifying the nature of the spam email. In such methods, none of the technological tools are used rather the users and companies are involved in terrifying the spammers. The proactive nature is an excellent feature of these solutions which can establish a reputation in the organisations that suffer from the problems of wastage of network bandwidth on downloading email spam. With proper awareness and devotion of the users, these solutions can give significant results.

#### 1.6.1.1 Recipient Revolts

If the user shows anger against the incoming spam e.g. an advertisement from legitimate companies, then these companies can be scared of loss of trust and reputation. Anticipation of a discontinued relationship with the user will force them to keep away from such kind of activities. Some advantages can be seen in this solution:

- This solution forces the Internet Service Providers (ISP’s) to formulate some useful policies.

- Legitimate companies will be scared of losing the trust and relationship of real customers.

- After popularisation of this method, a positive feedback can be obtained which will help to reduce the bulk emails.

Some of the disadvantages can also be seen:

- It will increase the burden for the ISP’s for identifying valid or invalid complaints.

- A proper authentication will be necessary so that only the right person is punished.
• It can be a cause of limited communication because spammers hide their identity that can force a user to block all the unknown ids.

1.6.1.2 Customer Revolts

Spam emails contain advertisements of different companies. This method indicates that companies to whom users disclose their personal data should be forced to reveal the purpose for which they have taken the data. They should also publish proper policies on the web pages where the purpose of obtaining the data should be mentioned. Some disadvantages of this solution are mentioned below:

• This solution may create hurdles in tackling false complaints.

• It’s hard to distinguish between valid and invalid complaints.

1.6.1.3 Vigilante Attack

This solution suggests some hard steps where spam addresses are to be treated angrily and punished with ‘mail bomb’ or ‘denial of service’ attacks. In this solution, some of the innocent users could be punished by being classified as a spammer. Some of the disadvantages of this solution are discussed below:

• Identification of the right person is important to ensure so that an innocent could not be punished.

• The results of this solution could be nasty or unethical.

1.6.1.4 Hiding the addresses

This solution suggests for having two email ids where first id is used for communication and other kept hidden. All the emails are delivered in the first id and manually inspected by the user. Further, all identified legitimate emails are forwarded to the second id. The second
email id is disclosed only to trusted persons so that only legitimate communication can be initiated.

Shortcomings of this method are:

- It’s hard to maintain two email ids.
- This method is just a simple thought for handling smaller volumes of mail, and cannot be seen a solution for stopping spam.

1.6.1.5 Legitimate contracts and limiting trial accounts

This solution works with an agreement between users and the Internet Service Providers. In this process, users sign an agreement before registering for new email address. By this method, ISPs can get the proper identity of users and easily distinguish valid and non-valid addresses. If the agreement is violated at any stage, appropriate action is taken against the user by blocking their id or giving proper punishment. This solution appears quite attractive but the main hurdle is the disclosure of identity of the users, hence it will not be accepted by some users.

1.6.2 Technical Solutions

Technological solutions are reactive in nature that means they work only when spam has already been delivered to the users’ account. Such methods do not force to spammers to stop spamming rather they fight to make the spammers’ job difficult. Nowadays, reactive solutions are more prevalent than proactive solutions in spam classification research. Some of the technological solutions are given below:
1.6.2.1 Domain Filters

In such filters, different programmes are written and configured so that they accept only those emails which are sent from specific domains. In this process, the email addresses, that have not included in the domains, will not be retrieved. Such filters have some disadvantages that are discussed below:

- In this method, Spammers can spoof valid emails that are present in legitimate domains.
- This method can narrow down communications by filtering even new valid ids.

1.6.2.2 Blacklisting

In this method, a database is created where bad senders’/spammers’ email ids and IP addresses are listed. This list is called a blacklist. Several users and organisations work collaboratively to maintain this list. When emails come from these addresses they are considered as spam and delivery blocked to in the users’ mailboxes. Some drawbacks of this method are discussed below:

- Maintaining the list of the spammers creates overhead costs.
- Since these databases are developed in a distributed manner, information retrieval and constant updation are complicated processes in the real time scenario.
- Database can be outdated soon because a user can change their email id anytime and hence difficult to recognise it quickly.

1.6.2.3 White-list Filters

In this method, the email program is configured to list all the addresses of good users and the emails coming from those addresses will only be allowed to pass. Emails from unknown
addresses will be kept in a separate folder. This technique makes the manual inspection for eliminating spam easy because users have to inspect the unknown addresses folder only. Some of the advantages of this technique are listed below:

- This method makes the inbox spam free because all the mails here would have come from known contacts only.
- For better results this method can be used in combination with other filtering techniques.

Some disadvantages can also be seen in this method:

- It’s hard to maintain a list of all the contacts because users can change their address frequently.
- Sometimes important emails that are delivered from the unknown source can be skipped.
- It also limits the range of communication.

1.6.2.4 Rule-based Methods

This solution works by developing a phrasal or word relation between instances and the corresponding classes. Several rules are formed to define this relationship. Sometimes several rules are combined to produce a concrete filtering solution. These rule-based methods work with the assignment of certain weights that identify a suitable class for an unknown instance. Some disadvantages of these methods are given below:

- These methods require human experts.
• Each rule based method becomes outdated frequently because spammers modify the spam words and nature of spam by identifying the rules.

1.7 **Motivation**

Although, legal action against delivery of spam has been initiated by legislation but a spammer can be attracted by the high profit and low market entry barrier (Zhang 2005). An email spam is associated with some negative externalities which result in some costs to the email service provider and the end user (Melville et al. 2006; Goodman, Cormack, and Heckerman, 2007). An email service provider handles billions of “abusive email” per day (MAAWG 2007).

Organisations’ concern is the productivity loss of employees due to spam, caused by limited attention to the email messages (Falkinger 2007). Hence, email spamming affects individuals (inside and outside of the organisations) and global scale economies. In response to this, it is necessary to take action to reduce the cost of spam at the technical, market and legislation level (OECD 2005; Joseph & Thevarajan 2008). Implementation of the market and legal action are long term processes and difficult for a IS manager. However, technical actions such as spam filtering approaches are successfully used by many organisations and users to reduce the number of spams in the mailboxes. Although, spam filtering is used by a number of users and organisations, no study has empirically addressed the issue of whether a spam filter reduces cost or not (Caliendo, Clement, Papis, & Scheel-Kopeinig 2008). The desired outcome of a spam filter cannot necessarily be seen because the net benefit of a spam filter depends on the cost incurred in the installation, time consumption during the updation and training process and for inspecting the results due to the threat of misclassification. Some studies have been done on cost statistics related to email spam and spam filtering systems.
According to the Commercial Internet Exchange association (CLX), 94% Internet Service Providers (ISPs) reported that spam irritates their subscribers. In addition, 79.5% said that unsolicited commercial emails (UCE) slow the system performance, 75.9% stated problems with operating cost, and 33.7% complained about the system outages.

Rao, Justin and Reiley, 2012, have done a study on cost associated with spam and spam filtering systems. According to them, a huge direct cost is incurred due to email spam. Statistics show that an end user suffers $30-$50 direct cost of spam per year. In organisations, $730 direct cost is incurred due to loss in productivity for each employee. Study estimates that US corporations bear $8.9 billion total cost per year due to spam. This study also explores the cost related to installation and maintenance of the spam filtering system that is observed as approximately $6.5 billion per year.

In addition, a major threat is seen by the phishing emails. A phishing email is used to steal the personal data like credit card numbers, passwords, account data, or other information of an user. It is also used to make vulnerable the systems employed by service providers. Spammers get huge benefit by such kind of mails. A report says that, in 2005, a spammer named Devid Levi made over $360,000 from 160 people using an eBay Phishing scam. In Dec. 2006, 28000 unique phishing emails have been reported which was almost double of the number indentified in 2005. This study is also observed that due to these phishing emails approximately $2 million cost incur for US Corporation in 2005. These corporations have also suffered with the additional losses due to customers fear (Harold & Cothern, 2007).

1.8 Research Objective:

This research covers the identified issues and challenges associated with a spam filter and address some major problems encountered with email spam and current spam filtering systems.
The foremost objective of this research is to develop an Spam Filtering System that cater to the basic need of Internet Service Providers, business enterprises and individuals. This system tackle the existing problems such as cost of installation, poor classification accuracy, complexity etc. related to spam and current spam filtering systems by improving the performance. The performance of this system will be captured by different metrics and done by:

- Improving the Classification accuracy (F-value);
- Reducing the False Positive Rate;
- Reducing the Training Time;
- Reducing the Testing Time.

This filtering system will provide convenient adaptability, and frequent updating and training capability to any user where specific emails examples could be used for training to construct a specific filtering model for this system.

1.8.1 Importance of F-value:

A popular and important measure for evaluating the performance of a classifier is performance accuracy which tells the willingness of the classifier to classify emails. In the literature, it has been identified that performance calculated by only accuracy metric is not enough because some time this parameter misjudges the correctness. To tackle with this problem, F-Value is involved that uses precision and recall value of classifier during the calculation of accuracy.

1.8.2 Importance of False Positive Rate:

Misclassification is very serious problem for a classifier especially when a legitimate (HAM) file misclassified as spam. When a spam file is misclassified as legitimate file, it can be easily observed and removed but if a legitimate file misclassify as spam, it would be a critical
problem because that file could be an important file. False Positive Rate (FP Rate) tells the rate of misclassified files which were legitimate. For a robust and accurate classifier, this value should be as low as possible.

1.8.3 Importance of Training and Testing Time:

The installation of a robust classifier can be decided not only by observing how accurate it would be but also how fast it would be. In this study, Training and Testing time are also employed to check the rapidness of the classifier during the training, and testing.

Remaining chapters of this thesis have been structured in this fashion. Chapter 2 presents the literature review of spam classifier. Chapter 3 demonstrate the Experimental design of this research. Chapter 4 shows the first study where various feature selection technique and feature subset selection technique are taken for comparison and identifying the best one. Chapter 5 presents a comparative study between various machine learning classifiers. Chapter 6 identifies the best pair of machine learning classifier and feature subset search techniques. Chapter 7 develops two different classifier models where first is developed by combining classifier technique and other by ensemble of classifiers approach. Chapter 8 is the last study of this research where all best classifiers model are taken to check their rapidness in terms of Training and Testing time. At the last chapter 9 concludes this research with many business implications, contribution of this research, future work and limitations.