Chapter 9. Conclusion, Future Work, Business Implications and Limitations

It has been observed in the literature that existing spam filtering systems suffer from several issues and challenges like cost of installation, training and testing time, poor performance accuracy, misclassification of positive instances and customization capability etc. hence the exact worth of these systems is difficult to capture. Most of the current spam filtering systems use the header based methods and are constructed by system approaches (Black-listing, White-listing, Grey-listing and Challenge Response systems etc.). Installation of such kind of systems is complex and costly. Only a few systems include content-based learning methods that provide excellent automation in spam filtering systems. The content-based filters have an important feature i.e. customization capability so that such filters can be installed by any organisations and internet service providers and trained by their own pre-existing corpuses.

The aim of this research was to develop a robust, fast, accurate, sensitive and customizable content-based spam filtering model that could cater the basic need of the organisations and the internet service providers. For achieving our objective, this research has conducted several comparative studies to identify the excellent existing algorithms for most informative feature selection and good machine learning. In addition, combining and ensemble of classifiers methods were also included in this research and some novel classification algorithms have been proposed.

At first, various Feature Selection Methods and Feature Subset Search Methods were experimented. The results of the analyses suggest that among all Feature Selection methods compared in this research, Relief F (RF) was found to be the best. On the other hand, Greedy

Stepwise search was found to be excellent Feature Subset Search Method. Further, various combinations of Machine Learning and Feature Subset Search methods were tested. In this study, Support Vector Machine and Random Forest with Greedy Stepwise Search method were found to be the best pair. In addition, Bayesian Classifier has again proved its worth with good accuracy and fast training and testing performance.

For reducing the learning and sampling errors, various ensemble-based approaches and combining classifiers approaches have also been tested in this research. In addition, a novel Enhanced Genetic Programming approach has been developed and compared with the other classifiers. This novel algorithm has proved to be the best of this study in terms of classification accuracy and False Positive Rate.

Finally, we have successfully achieved our objective by developing two different models that have been tested and found to be excellent in tackling the issues of current spam filtering systems. Both models have shown some trade-off between classification accuracy and Training Time. The first model, developed from Enhanced Genetic Programming Classifier with Greedy Feature Subset Search Method, was found robust, fast (in Testing), accurate and sensitive with less false positive rate but at the Training Time was high compared to other models. On the other hand, the second model was built from Bayesian Classifier with Greedy Stepwise Feature Search method, was found robust, rapid, sensitive towards accuracy but the accuracy was found less than the first model proposed.

9.1 Business implication:

The business implication of this research consists of the reduction in cost incurred due to spam/unsolicited bulk email. Email is a fundamental necessity to share information within a number of units of the organisations to be competitive with the business rivals. In addition, it

is continually a hurdle for internet service providers to provide best emailing services to their customers. Although, the organisations and the internet service providers are continuously adopting novel spam filtering approaches to reduce the amount of unwanted emails, but the desired effect could not be significantly seen due to the cost of installation, vast training time, customizable ability and the threat of misclassification of important emails. This Research deals with all the issues and challenges faced by Internet Service Providers and Organisations. In this research, the proposed models have not only provided excellent performance accuracy, sensitivity with low false positive rate, customizable capability but also worked on reducing the cost of installation, training and testing time.

9.2 Contribution of this research:

In this research, some comparative studies between feature selection, feature subset search and machine learning classifiers have been performed. In addition to this, a novel ensemble based and a combining classifier with committee selection method has been proposed. This research contributes in the research by Observing and validating RF (Relief F) as a best feature selection and Greedy stepwise search as best feature subset search method. On the other hand, SVM (Support Vector Machine) with Greedy Stepwise search method was the best combination of machine learning and feature subset search method. For improving the learning of the SVM, NP (Normalised Polynomial) kernel has been observed and validated as best kernel. For reducing the sampling error and improving the learning of classifiers, adaboost has been predicted to be best boosting algorithms. In addition, this research has proposed and validated two novel machine learning classifiers. The first was combining classifier with committee selection and second was EGP (Enhanced Genetic Programming) classifier. This research ends up by proposing two models; one is EGP with Greedy Stepwise search and other is Bayesian with Greedy Stepwise search.

9.3 Future Work:

Researchers and Developers of the spam classification domain can get benefited from this research by extending this work. This research has provided a basic building block of robust, fast, accurate and customizable filtering systems. In addition, different feature subset search methods can also be employed in the proposed models for making them more robust. The proposed models have been tested only for classifying the email spams however they can also be used for other applications of data and text mining. In the literature, only body part of the emails was found suitable for content-based filtering that was also the choice of this research. In the continuation, body, header and the combination of them (body + Header) can be further included and compared with each other to identify most suitable part of the emails for proposed models. Although, different corpuses have been incorporated for testing the customizable capability of the proposed models, the consistent performance can be further checked by including other complex corpuses.

9.4 Limitations

This research is focussing on the classification of those email spams that are written in the English language only because English a universally accepted language that is widely used to write emails. None other languages have been included in this study. As discussed in chapter 2 (section 2.5.4), body part of email files is more suitable for content-based filtering. This method works with the content (words/features) of the email data that is found in the body part; hence, this research has employed only body part of the emails. A most critical category of spam is image spam where the content of spam is imbibed into the image file discussed in chapter 1 (section 1.3.4). It's hard to capture this content. Some other content recognition methods such as Optical Character Recognition (OCR) and Image Processing etc. are employed for extracting the content of image files and considered as another area of the

research so that Image spam has not been included in this research. Only emails messages have been incorporated into this research. None other messages like Short Message Service (SMS) or Multi-Media Messaging Service (MMS) were a part of this study.