Abstract

In today’s automated world, to be competitive and sustainable in business, information sharing within the units of the organisation is necessary. Email is an essential and useful tool of rapid and cheap communication. It is now a popular medium to connect people with each other. On the other hand, Spam (also known as unsolicited bulk email) is a challenge for the organisations and the researcher because its size is increasing day by day. This rapid growth causes serious hitches, such as unnecessary filling of users’ mailboxes, engulfing some important emails, consuming storage space and bandwidth as well as too much time consumption in sorting them.

Various email spam filtering systems have been implemented by different organisations and the Internet Service Providers (ISPs) but none is found perfect. The existing systems suffer from a number of problems (like installation cost with no guaranteed positive outcomes, high false positive rate, training and testing time etc.) which are directly and indirectly reflected in the cost incurred due to spam and inadequate spam filters.

Recently, content-based filters are gaining popularity in the spam classification domain. Various machine learning algorithms are used in this method to classify emails. This method works with the content of the documents that are extracted to distinguish unsolicited email and legitimate emails. A critical issue has been identified in this method i.e. misclassification of the legitimate emails. Sometimes legitimate emails carry valuable information for a user, and if such good emails misclassify as spam, it will create serious problems. To tackle this issue, False Positive rate is used that evaluate the rate of misclassified legitimate emails. For a robust and sensitive classifier, the False positive rate should be as low as possible (ideally zero).

The need of the researchers and developers is to construct such filtering system that can be acceptable for any organisation and the Internet service providers. Content-based spam filters are popular due to their customization capability where, pre-existing email content can be used to train such filters. The emailing pattern of each organisation differs hence they need to install such filtering system which can be trained by particular words/features of their own pre-existing corpus so that the performance accuracy can be maximised and the False Positive Rate can be minimised.

This research proposes a content-based email spam filtering system with a set of novel algorithms along with some existing algorithms at the levels of feature selection as well as learning. Various studies have been done in different chapters for achieving this goal where several comparative studies between Feature Selection Methods, Feature Subset Search Methods, and Machine Learning Classifiers have been performed, together with some novel approaches proposed. All the tests have been done on popular publically available corpuses. According to the analysis, Relief F is predicted to be the best Feature Selection Method amongst other methods compared whereas Greedy Stepwise Search was found to be the best
feature subset search method in terms of good performance accuracy and low False Positive Rate.

Further, among Machine Learning Classifiers, Support Vector Machine and Random Forest are identified to be excellent whereas Bayesian classifier proved its worth in this domain. Support Vector Machine together with Greedy Stepwise Feature Search was the best pair of the Machine Learning and Feature Subset Search Method. Results have shown that this pair was excellent in terms of good performance accuracy and low FP rate amongst other pairs.

For Improving the Performance of classifiers, various combining and ensemble-based techniques were also included in this research. For combining method, the best combination of classifiers (boosted Probabilistic classifiers and Support Vector Machine) has been used to construct a combined classifier with committee Selection Mechanisms.

A novel Ensemble based Approach with the Genetic Programming Classifier has been developed and tested together with the best classifiers. It has been identified that this Enhanced Genetic Programming (EGP) Classifier is promising from all the performance dimensions.

Finally, this research has achieved the objective that was proposed initially by developing two different filtering models that have a trade-off between Accuracy and Training Time. First model incorporates Enhanced Genetic Programming Classifier with Greedy Stepwise feature search method and proves to be robust, fast (in Testing Time), accurate, and sensitive with less false positive rate but with high Training Time compared to the other models. Second model, that has been developed with Boosted Bayesian classifier and Greedy Subset feature search method, is robust, fast (both in Training and Testing), and sensitive with less false positive rate but classification accuracy is less than the first model proposed. In addition, the proposed filtering models satisfactorily tackle the issue of customization so that organisations can install these models to train according to their need.

The Proposed Models will cater the need of Organisations and Internet Service Providers (ISPs) where after installation they can minimize the cost related to spam and spam filtering system.

This research has contributed in the literature of spam classification by the following way:

1. Observed and validated RF (Relief F) as a best feature selection method.
2. Observed and validated Greedy Stepwise search as a best feature subset search method.
4. Observed and validated AdaBoost as best boosting method.
5. Observed and validated SVM (Support Vector Machine) as a best machine learning classifier.
6. Observed and validated SVM and Greedy Stepwise search as a best combination of machine learning and feature subset search methods.

7. Developed and validated a novel combining classifier with committee selection method.

8. Developed and validated a novel EGP (Enhanced Genetic Programming) machine learning classifier.

9. Two new models have been proposed:
   - EGP with Greedy Stepwise search.
   - Bayesian with Greedy Stepwise search.

**Keywords:** Spam Classification, Machine Learning Classifiers, Feature Selection Methods, Feature Subset Search Methods, Ensemble of Classifiers, Enhanced Genetic Programming Classifier, Combining Classifiers, Performance Accuracy, F-Value, False Positive Rate, Training Time, Testing Time.