**LITERATURE SURVEY**

# 1) Machine Learning based Spam E-Mail Detection

# AUTHORS: Priti Sharma, Uma Bhardwaj

Spam email is one of the biggest issues in the world of internet. Spam emails not only influence the organisations financially but also exasperate the individual email user. This paper aims to propose a machine learning based hybrid bagging approach by implementing the two machine learning algorithms: Naïve Bayes and J48 (decision tree) for the spam email detection. In this process, dataset is divided into different sets and given as input to each algorithm. Total three experiments are performed and the results obtained are compared in terms of precision, recall, accuracy, f-measure, true negative rate, false positive rate and false negative rate. The two experiments are performed using individual Naïve Bayes & J48 algorithms. Third experiment is the proposed SMD system implemented using hybrid bagged approach. The overall accuracy of 87.5% achieved by the hybrid bagged approach based SMD system.

# 2) Email Header Feature Extraction using Adaptive and Collaborative approach for Email Classification

# AUTHORS: Amandeep Singh Rajput, J S Sohal, Vijay Athavale

# Email Header is footprint of an Email that can be used to examine an Email as HAM or SPAM. Email classification in this research is done on the basis of header features thus by keeping the content privacy of the sender intact [1]. Header features are , email header fields like sender, to, cc, bcc, subject. This research tries to improve the accuracy of the classification by extracting more number of header features. Email Subject is further deeply examined for objectionable keywords for rule matching and rule generation. In our study, we implement an adaptive and collaborative approach by using machine learning and cluster computing for fast classification of Emails as SPAM or HAM. Adaptive approach is to generate new rules for classification and cluster approach is to use parallel computing power for increasing computing speed. New rules are only generated if features extracted from email header do not match the existing rules. Spam Assassin [2][3] is the main dataset used for testing. Collaborative approach creates a parallel environment where multiple antispam methods and divided test corpora are used as input. The false positive and false negative percentage are recorded and accuracy is calculated. Weka Data Mining Software is used to apply the anti-spam methods.

# 3) Effect of Header-based Features on Accuracy of Classifiers for Spam Email Classification

**AUTHORS** **: Dr. Priti Kulkarni1 , Prof & Dr. Jatinderkumar R. Saini2**

Emails are an integral part of communication in today’s world. But Spam emails are a hindrance, leading to reduction in efficiency, security threats and wastage of bandwidth. Hence, they need to be filtered at the first filtering station, so that employees are spared the drudgery of handling them. Most of the earlier approaches are mainly focused on building content-based filters using body of an email message. Use of selected header features to filter spam, is a better strategy, which was initiated by few researchers. In this context, our research intends to find out minimum number of features required to classify spam and ham emails. A set of experiments was conducted with three datasets and five Feature Selection techniques namely Chi-square, Correlation, Relief Feature Selection, Information Gain, and Wrapper. Five-classification algorithms-Naïve Bayes, Decision Tree, NBTree, Random Forest and Support Vector Machine were used. In most of the approaches, a trade-off exists between improper filtering and number of features. Hence arriving at an optimum set of features is a challenge. Our results show that in order to achieve the objective of satisfactory filtering, minimum 5 and maximum 14 features are required. Keywords.

**4) Machine learning for email spam filtering: review, approaches and open research problems**

**AUTHORS :** **[Emmanuel GbengaDada](https://www.sciencedirect.com/science/article/pii/S2405844018353404" \l "!)[a](https://www.sciencedirect.com/science/article/pii/S2405844018353404" \l "!)****,**[**Joseph StephenBassi**](https://www.sciencedirect.com/science/article/pii/S2405844018353404#!)

The upsurge in the volume of unwanted emails called spam has created an intense need for the development of more dependable and robust antispam filters. [Machine learning methods](https://www.sciencedirect.com/topics/engineering/machine-learning-method) of recent are being used to successfully detect and filter spam emails. We present a [systematic review](https://www.sciencedirect.com/topics/psychology/systematic-review) of some of the popular machine learning based email spam filtering approaches. Our review covers survey of the important concepts, attempts, efficiency, and the research trend in spam filtering. The preliminary discussion in the study background examines the applications of [machine learning techniques](https://www.sciencedirect.com/topics/computer-science/machine-learning-technique) to the email spam filtering process of the leading [internet service providers](https://www.sciencedirect.com/topics/computer-science/internet-service-provider) (ISPs) like Gmail, Yahoo and [Outlook](https://www.sciencedirect.com/topics/nursing-and-health-professions/angiographic-catheter) emails spam filters. Discussion on general email spam filtering process, and the various efforts by different researchers in combating spam through the use machine learning techniques was done. Our review compares the strengths and drawbacks of existing [machine learning approaches](https://www.sciencedirect.com/topics/computer-science/machine-learning-approach) and the open research problems in spam filtering. We recommended deep leaning and deep adversarial learning as the future techniques that can effectively handle the menace of spam emails..

**5) Efficient email classification approach based on semantic methods**

# AUTHORS: [Eman M.Bahgat](https://www.sciencedirect.com/science/article/pii/S2090447918300455" \l "!), [SherineRady](https://www.sciencedirect.com/science/article/pii/S2090447918300455#!)

# Emails have become one of the major applications in daily life. The continuous growth in the number of email users has led to a massive increase of unsolicited emails, which are also known as spam emails. Managing and classifying this huge number of emails is an important challenge. Most of the approaches introduced to solve this problem handled the high dimensionality of emails by using syntactic feature selection. In this paper, an efficient email filtering approach based on semantic methods is addressed. The proposed approach employs the WordNet ontology and applies different semantic based methods and similarity measures for reducing the huge number of extracted textual features, and hence the space and time complexities are reduced. Moreover, to get the minimal optimal features’ set, feature dimensionality reduction has been integrated using feature selection techniques such as the [Principal Component Analysis](https://www.sciencedirect.com/topics/engineering/principal-component-analysis) (PCA) and the Correlation Feature Selection (CFS). Experimental results on the standard benchmark Enron Dataset showed that the proposed semantic filtering approach combined with the feature selection achieves high computational performance at high space and time reduction rates. A comparative study for several [classification algorithms](https://www.sciencedirect.com/topics/engineering/classification-algorithm) indicated that the Logistic Regression achieves the highest accuracy compared to Naïve Bayes, [Support Vector Machine](https://www.sciencedirect.com/topics/engineering/support-vector-machine), J48, Random Forest, and [radial basis function networks](https://www.sciencedirect.com/topics/engineering/radial-basis-function-network). By integrating the CFS feature selection technique, the average recorded accuracy for the all used algorithms is above 90%, with more than 90% feature reduction. Besides, the conducted experiments showed that the proposed work has a highly significant performance with higher accuracy and less time compared to other related works.