**In t r o d u c t io n**

The Internet as a “network of networks” has expanded the possibilities of communication and placement of content. Email system is one of the most effective and commonly used sources of communication [1]. Unfortunately, the continuous rise of email users has led to a massive increase of spam emails. Spam emails are usually sent in bulk and do not target individual recipients. Whether it is commercial in nature or not, spam emails can cause serious problems in electronic

communication. Spam emails produce huge amount of unsolicited data and thus affect the network bandwidth and storage capacity. Due to the large number of spam emails to users of email services it is difficult to distinguish useful from unsolicited emails. Thus, managing and filtering emails is an important challenge. The filtering purpose is to detect and isolate spam emails. There are two main approaches for spam detection. The first approach is based on email header analysis and the second one is based on email body analysis. Spam filters usually combine those two approaches. Email header include fields like From, To, Subject, CC (Carbon Copy), BCC (Blind Carbon Copy) which almost reveals the nature of the email. The recent studies have shown that the information provided by email header is quite important [2], [3]. Content based filtering relies on the assumption that the body content of spam email is different than the legitimate or ham mail. In recent years, a number of Machine Learning (ML) and data mining techniques have been engaged in order to classify email messages based on its content. Classification methods, such as Naive Bayes, Support Vector Machine, Decision Tree, Random Forest and Neural Networks are commonly used to develop efficient email classifier [4]. For the most classification problems the process of feature extraction and selection from email body is very important. In this paper, semantic properties of email content are used for feature selection and reduction. In order to efficiently detect spam emails various preprocessing steps need to be done, such as stop words removal, stemming, term frequency [5], [6], [7]. The aim is to preserve the most important features and to reduce computations demand. After feature selection, the ID3 algorithm is used to generate a decision tree that categorizes emails as spam or ham [8], [9]. The proposed approach is evaluated using accuracy, precision and recall. The performance of proposed system is measured against the size of dataset and feature size. This paper is organized as follows. Section II explains proposed approach for spam detection in detail. Section III summarizes the results while Section IV gives the conclusion.