Email Classification Using Machine Learning Algorithms

Anju Radhakrishnan [#1], Vaidhehi V [*2]
[#] Department of Computer Science, Christ University, Bengaluru, India
[1] anju.radhakrishnan@cs.christuniversity.in
[2] vaidhehi.v@christuniversity.in

***Abstract*— Email has become one of the frequently used forms of communication. Everyone has at least one email account. Inflow of spam messages is a major problem faced by email users. Currently there are many spam filtering techniques. As the spam filtering techniques came up, spammers improved their methods of spamming. Thus, an effective spam filtering technique is the timely requirement. In this paper email classification is done using machine learning algorithms. Two of the important algorithms namely, Naïve Bayes and J48 Decision Tree are tested for their efficiency in classifying emails as spam or ham. The experiment focused on classification in combination with pre-processing techniques and concepts of text categorization. The dataset used is Enron Corpus. TF-IDF value is used as the weight score of text. The classifiers are also tested for different feature size. The test results show that J48 is more accurate in classifying emails as spam or ham with a minimum feature size and classification time.**

Keyword - Spam, Machine learning algorithm, Naïve Bayes, J48 Decision Tree, Pre-processing, Enron, TF-IDF, Feature size.

I. INTRODUCTION

Email is one of the most efficient and effective mode of communicating with one another. Today a serious problem for web users and web services is caused by inflow of large number of spam emails. Spam mails are called the unwanted mails or unsolicited mails or bad emails which user receives without any prior information of the sender. Spam mails are used for spreading viruses, advertisements, for fraud in banking and for phishing. So, it can cause serious problems for internet users like loss of data, waste of time and energy of users, loading traffic on the network. To avoid irrelevant mails, need of effective mail filtering method is a timely requirement.

Filtering is nothing but the method of arranging mails, such as removing spam mails, deleting viruses, and only non-spam mails can flow in. Existing spam filtering techniques use classification. Classification is the technique of data mining. Data mining is defined as discovering useful knowledge from large data. Classification is the process of finding a model that describes and distinguishes different classes or concepts of data. The models are derived based on the analysis of set of objects of different classes for which the class label is known. Classification is a type of data analysis that extracts models describing important data classes or concepts. Classification mainly consists of two steps. First is the learning step: where a classification model is constructed and second is the classification step: in this step the extracted model is used to predict the class labels for new data or unknown data depending on the learning step.

Machine learning algorithms are used for classification of objects of different classes. Such algorithms have proved to be efficient in classifying emails as spam or ham. This research work has used two main machine learning algorithms namely, Naïve Bayes and J48 Decision Tree. Enron Email Corpus has been used for experiment. It is one of the publicly available large datasets of email. The tool used for the experiment setup is weka.

Over the last few decades' digital files and databases are used mostly for storing the data. At the same time the users of this same data are expecting to find more sophisticated information and hidden patterns from them. Text mining is a form of data mining which is used to uncover the hidden information in textual data. A major application of text mining is web text mining which is used for filtering emails as spam or ham. Instead of the simple word counts the TF-IDF value of words has been used for analysing the data objects. TF-IDF is short form for the term frequency- inverse document frequency. It is considered as a successful measure in text summarization and classification. TF-IDF is often used in text mining as a weighting score of text.

Classification accuracy is found more when the TF-IDF value is used compared to simple word counts. The inverse document frequency shows how much information the text provides, whether the text is common or rare across all documents provided. The results of the experiment show that J48 Decision Tree has a higher accuracy of email classification as spam or ham compared to Naïve Bayes for a lesser number of feature size and classification time.

II. RELATED WORK

Spam emails referred to as junk emails or unsolicited bulk emails. Spam emails have become one of the biggest threat for today's internet users. Spam emails cause problems like occupying large space in inbox, spreading viruses, causing loss of valuable information etc. Many techniques have been proposed to filter such emails. A method has been proposed for identification and screening of spam emails by using Rapid Miner data mining tool. Major emphasis is on pre-processing and the importance of pre-processing techniques in text mining. Different data mining pre-processing techniques have been used. Classification algorithms are used over the taken dataset after pre-processing. The classifiers are evaluated based on different parameters like accuracy, precision, execution time etc.[1]. Different Decision Tree classifiers are analysed for their accuracy in classifying emails through data mining approach. Naïve Bayes Tree classifier, J48 Decision Tree classifier and LMT were analysed. In terms of performance and accuracy level LMT showed best results[2].

An algorithm is proposed using Naïve Bayesian theorem. The algorithm classified emails as spam or ham. The classification of the mails were based on the email body content[3]. Different classification techniques used in email classification like SVM, K- means clustering, vector space model etc. are discussed based on the features and their limitations. The results showed that unsupervised learning is ignored in most of the cases and data mining techniques proved to be effective[4]. A method for clustering spam messages using genetic algorithm and k-nearest neighbour algorithm are proposed. Showed the efficiency of clustering method in grouping emails as spam or legitimate[5].

Different spam detection approaches are discussed. Spam detection techniques starts with non-machine learning to machine learning. Based on false positive and false negative rates different approaches for classifying emails are studied. The results showed no method provides 100% efficiency[6]. Text clustering method used for spam detection based on vector space model is proposed. Data is used represented using vector space model. An algorithm based on k-means and BIRCH algorithm is presented. K-means proved to be best for smaller datasets[7]. What is spam and various problems caused by spam were discussed. Spam filtering trends and techniques are analysed. The techniques discussed are deployed on receiver side[8].

A method for spam identification based on characters-word is proposed. The classifier used for the approach is multi-neural networks. ASCII value of the word characters are used as the weighting factor for attributes. High false positive and low true negative rates are achieved. Good or bad words are identified using the approach. Messages are pre-processed before giving to the classifier[9]. A method using data reduction for email classification has been proposed. Instance Selection Method is used to avoid irrelevant information from the dataset before classifying. The results showed good accuracy in classification with reduction of false positive instances. Cluster classifiers been used for data reduction from training model[10].

A new method of classifying emails using pattern discovery technique is used. Naïve Bayesian algorithm in combination with pattern discovery proved to be best in identifying spam messages. The proposed approach gave best results comparing to other methods such as data mining methods. The proposed approach used text mining[11]. The traditional and learning based approaches in spam detection are studied. An overview of existing spam filtering techniques is given. Learning based algorithms proved to exceed traditional ones because of number of qualities[12].

Different classification algorithms are analysed using weka tool in classification of emails as spam or ham. Spambase dataset from UCI repository is used for the experiment. The classifiers used are Naïve Bayes, Logistic, PART and J48. The results showed that Logistic classifier algorithm performs best for spambase dataset[13]. Various problems associated with spam and spam filtering methods are analysed. Both traditional and learning based methods are studied. The results showed that no methods provide 0% false positive and 0% false negative rates[14].

Enron Email Corpus is introduced as a new test bed in spam filtering and text learning research. Explained about the suitability of Enron Corpus with respect to email folder prediction. The Enron Corpus of email was evaluated using another email dataset. The experiment showed that the Enron corpus is a large standard test dataset that could be very much valuable[15]. Data classification technique is used for spam filtering. In content based detection, classification and semi supervised learning are the frequently used approaches. Spam filtering is an application of content based detection. The experiments showed that content based techniques outperforms rule based techniques[16].

A technique of email mining using k-mean clustering algorithm and weka tool is proposed. Text converter used to process the whole data into .csv file format from .eml format. The implementation focused on filtering the email addresses from which maximum emails are generated. The implemented method helps the users to take a back control of their mailboxes[17]. Traditional approaches of spam filtering like blacklisting/whitelisting, keyword matching, content based detection proved to be no more efficient in spam filtering. This is because of the improved techniques by spammers. A method of spam filtering using supervised machine learning algorithms is proposed. The work used C4.5 Decision Tree, Naïve Bayes Classifier and multilayer perceptron. Personal mail dataset is used for the experiment. The classifiers were compared based on training time, correctly classified instances, prediction accuracy and false positive. Multilayer perceptron found to be outperforming other classifiers[18].

A method has been proposed based on Natural Language Processing which will enhance online security. It is a stepwise method which analyses the sender as well as the content of the email. It used traditional approach of spam filtering such as blacklisting and keyword matching. The method introduced a threshold counter which helps in reducing the web server congestion and improving spam filtering efficiency[19].

Many techniques have been proposed in the field of spam filtering. The studies show that machine learning methods have proved to perform better than other traditional approaches. But no techniques provide 0% false positive and 0% false negative rates. As the spammers improve their techniques of spamming, the existing spam filtering technologies also need to be improved. Several researchers have analysed the efficiency of various machine learning algorithms in spam email filtering approaches. The papers [2], [3], [13] and [18] evaluated different classifiers in correctly classifying spam mails. The use of Enron corpus in researches regarding spam filtering was discussed in paper [15]. However, the research is lacking the minimum feature size required for efficient filtering and the importance of pre-processing. This research work has used machine learning algorithms in combination with extensive pre-processing techniques and concepts of text categorisation. The experiment used Enron Corpus, which is the largest email dataset that is publicly available. TF-IDF value is emphasised in this experiment. The feature size for different classifiers that will give highest accuracy in classification is also tested.

III.METHODOLOGY

There are many techniques that are proved to filter spam emails. But the accuracy and efficiency of such spam email filtering techniques is still a question mark. Till now no such filtering techniques is proved to provide 100% accuracy in classification of emails as spam or ham. In this paper the proposed approach of spam filtering uses concepts of text mining for classification. The proposed approach of spam email classification consists of two main steps, pre-processing and classification. Pre-processing includes stop words removal, case transformation, stemming and tokenizing. Every email consists of message header and message body. The message body consists of text data. In this paper, only the message body content and subject line has been focused. Two main data mining algorithms have been used namely Naïve Bayesian and J48 Decision Tree.

Enron Email dataset is used in the implementation. Enron corpus is a publicly available email dataset. It consists of two folders namely spam and ham. The email dataset has 3672 ham emails and 1500 spam emails. Both folders contain .txt files with the message subject and body content. The email dataset is converted into ARFF format using the TextDirectoryLoader. The converted dataset is again transformed using StringToWord vector filter. The dataset is then pre-processed by the following steps, lowercasing, stop words removal, stemming and tokenizing. The training dataset is then given to the classifier. In this paper the dataset is tested for different number of attributes or words in emails contained in the Enron dataset. The accuracy of the two classifiers Naïve Bayes and J48 Decision Tree in classification of emails for different number of words and their classification time are tested.

Pre-processing is the primary step in data mining. In real world, most of the data are not complete, contains incorrect values, missing values etc. The accuracy of classification or mining depends on the data being used. So, the first and foremost step to be performed before mining task is to pre-process the data. Since the Enron email data set contains two folders with number of emails as .txt files, is transformed into a data mining compatible format. The dataset is converted into ARFF (Attribute Relation File Format). In this present work, Weka tool, has been used to evaluate the test results. Waikato Environment for Knowledge Analysis is a popular machine learning software written in Java. Weka supports several standard data mining tasks like pre-processing, classification, clustering, feature selection, regression and visualization.

Weka provides certain filtering algorithms that can be used to transform the data from numeric to discrete etc. After converting the dataset into ARFF, the data is transformed from string to word by StringToWordVector filter. Stop words are removed from the dataset which are irrelevant for the data mining task. All the data are lowercased and stemming is performed on the dataset. Snowball stemmer is used to do the so-called step. It is a small string processing language designed for use in information retrieval. All the words in the dataset are reduced to their stem form avoiding unnecessary information. Tokenizing is another important step in pre-processing. Alphabetic tokenizer is used to tokenize the words. It considers only words with alphabets. The weighting factor for each of the attribute or word is TF-IDF. It is the Term Frequency-Inverse Document Frequency which is calculated by multiplying the number of occurrences of a token in a document with logarithm of the quotient of total number of documents and number of documents with the specified token.

Pre-processed data is given to the classifier. The efficiency of the classifier is tested for different number of attributes or words. In this present work, Naïve Bayes and J48 classifiers are tested for their efficiency in classifying emails as spam or ham. The number of attributes is chosen at a scale of 100. The accuracy of both classifiers in correctly classified and incorrectly classified are recorded along with their classifying time.

IV. RESULTS

In the present work of spam classification, a combination of pre-processing techniques and classification are used. Text mining is a variation of data mining. The textual information in the email such as subject message body are analysed.

The efficiency of two machine learning algorithms Naïve Bayes and J48 Decision Tree are tested for different feature size. The dataset used for the experiment is Enron Corpus. The efficiency of J48 classifier is given in table 1 and the efficiency of Naïve Bayes classifier is given in table 2.

TABLE I. Efficiency of j48 Decision Tree Classifier for different feature size

No. of Attributes	Correctly Classified Percentage	Incorrectly Classified Percentage	Time Taken (Sec)
200	95.8237	4.1763	0.59
300	96.4617	3.5383	0.45
400	**96.5971**	3.4029	0.06
500	96.3844	3.6156	0.08
600	95.7436	4.2537	0.06
700	96.249	3.751	0.11
800	95.6883	4.3117	0.11
900	96.2104	3.7896	0.1
1000	96.307	3.693	0.13
1100	96.1137	3.8863	0.4
1200	95.901	4.099	0.15
1300	96.307	3.693	0.16
1400	96.2104	3.7896	0.19
1500	96.2104	3.7896	0.25
1600	96.2104	3.7896	0.2
1700	96.2104	3.7896	0.08
1800	96.133	3.867	0.06
1900	96.133	3.867	0.09
2000	96.133	3.867	0.07
2100	96.133	3.867	0.09
2200	96.133	3.867	0.09
2300	96.133	3.867	0.08
2400	96.133	3.867	0.07
2500	96.133	3.867	0.09
2600	96.133	3.867	0.09
2700	96.133	3.867	0.1
2800	96.133	3.867	0.09
2900	96.133	3.867	0.09
3000	96.133	3.867	0.1
3100	96.133	3.867	0.1
3200	96.133	3.867	0.11

TABLE II. Efficiency of Naïve Bayes Classifier for different feature size

No. of Attributes	Correctly Classified Percentage	Incorrectly Classified Percentage	Time Taken (Sec)
200	91.512	8.488	1.16
300	92.3821	7.6179	0.93
400	92.5367	7.4633	1.12
500	92.8848	7.1152	1.38
600	92.6527	7.3473	1.7
700	92.9814	7.0186	2.3
800	93.2521	6.7479	2.44
900	93.3488	6.6512	2.73
1000	93.9675	6.0325	3.08
1100	93.7935	6.2065	3.53
1200	94.0062	5.9938	3.93
1300	94.0642	5.9358	4.46
1400	94.2575	5.7425	4.71
1500	94.2769	5.7231	4.96
1600	94.2189	5.7811	5.07
1700	94.2382	5.7618	6.06
1800	94.2769	5.7231	5.92
1900	94.3542	5.6458	6.41
2000	94.2962	5.7038	6.84
2100	94.3155	5.6845	7.16
2200	94.3155	5.6845	7.57
2300	94.2962	5.7038	8.19
2400	93.9869	6.0131	10.55
2500	94.2962	5.7038	10.84
2600	94.3155	5.6845	34.34
2700	94.3735	5.6265	34.39
2800	**94.4122**	5.5878	36.33
2900	**94.4122**	5.5878	36.85
3000	**94.4122**	5.5878	38.53
3100	**94.4122**	5.5878	40.81
3200	**94.4122**	5.5878	40.52

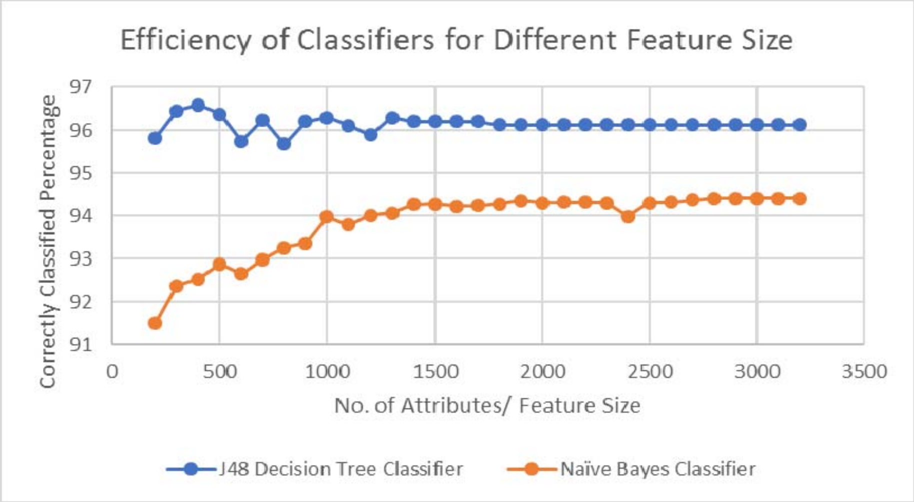

Fig. 1. Efficiency of Classifiers for Different Feature Size

V. CONCLUSION AND FUTURE ENHANCEMENT

The present experiment of email classification using machine learning algorithms showed that J48 Decision Tree Classifier is more efficient than the Naïve Bayes classifier for the dataset Enron Corpus. It gives an accuracy of 96.5971% in classifying the emails with a feature size of 400 attributes within a short span of time 0.06 seconds.

Further enhancements can be done to improve the efficiency of the classifier. This experiment has not reached the maximum efficiency that is 0% false positive and 0% false negative. In future, the work can be modified by using combination of classifiers in addition with pre-processing techniques.

ACKNOWLEDGEMENT

This research was supported by the Department of Computer Science, Christ University. We thank our professors who provided insight and expertise that greatly assisted the research. We would also like to show our gratitude to Prof. Joy Paulose for sharing his pearls of wisdom with us during the course of this research.

REFERENCES

[1] Proceedings of The International Conference on Innovations in Intelligent Systems and Computing Technologies, Philippines, “Text Mining Approach to Detect Spam in Emails,” no. February, 2016.
[2] S. Chakraborty and B. Mondal, “Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis,” Int. J. Comput. Appl., vol. 47, no. 16, pp. 26–31, 2012.
[3] C. Science and S. Engineering, “Effective Email Classification for Spam and Non-Spam,” vol. 4, no. 6, pp. 273–278, 2014.
[4] A. R. On and D. Glaucoma, “a Review on Different spam Detection,” vol. 11, no. 6, pp. 2–7, 2015.
[5] R. M. Alguliev, R. M. Aliguliyev, and S. A. Nazirova, “Classification of Textual E-Mail Spam Using Data Mining Techniques,” vol. 2011, 2011.
[6] M. T. Scholar and C. Science, “A Review on different Spam Detection Techniques.”
[7] M. Basavaraju and R. Prabhakar, “A Novel Method of Spam Mail Detection using Text Based Clustering Approach,” vol. 5, no. 4, pp. 15–25, 2010.
[8] S. Geerthik and T. P. Anish, “Filtering Spam : Current Trends and Techniques,” vol. 3, no. 8, pp. 208–223, 2013.
[9] A. Nosseir, K. Nagati, and I. Taj-eddin, “Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks,” no. November 2016, 2013.
[10] “Deakin Research Online,” pp. 1–5, 2010.
[11] “Spam Detection Using Baysian with Pattren Discovery,” no. 3, pp. 139–143, 2013.
[12] S. Nazirova, “Survey on Spam Filtering Techniques,” vol. 2011, no. August, pp. 153–160, 2011.
[13] S. S. Shinde and P. R. Patil, “International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS) Improving spam mail filtering using classification algorithms with discretization Filter,” pp. 82–87, 2014.
[14] V. Christina, P. S. G. R. K. College, and P. S. G. R. K. College, “A Study on Email Spam Filtering Techniques,” vol. 12, no. 1, pp. 7–9, 2010.
[15] B. Klimt and Y. Yang, “The Enron Corpus : A New Dataset for Email Classification Research.”
[16] M. E. Yitagesu and P. M. Tijare, “Email Classification using Classification Method,” vol. 32, no. 3, pp. 142–145, 2016.
[17] A. T. Sharma and M. I. Technology, “Analysis of Email Fraud Detection Using WEKA Tool,” vol. 10, no. 4, pp. 203–207, 2014.
[18] R. Giyanani and M. Desai, “Spam Detection using Natural Language Processing 1 1,” vol. 16, no. 5, pp. 116–119, 2014.
[19] V. Christina, S. Karpagavalli, and G. Suganya, “Email Spam Filtering using Supervised Machine Learning Techniques,” vol. 2, no. 9, pp. 3126–3129, 2010.

AUTHOR PROFILE

Ms. Anju Radhakrishnan currently pursuing her M.Sc. Computer Science from Christ University, Bengaluru India.

Mrs. Vaidhehi V Associate Professor at Christ University, Bengaluru India