International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET218323

Content Based E-Mail Classification

[1]Er.Sonal Chakole, [2]Sarita Padole, [3]Apurva Kamble, [4]Vandana Wadekar, [5]Ankit Dhande

[1]Assistant Professor, [2-5]BE Scholar, Department of Computer Science and Engineering, Priyadarshini J. L. College of Engineering, Nagpur, Maharashtra, India

Article Info
Volume 8, Issue 3
Page Number: 141-144

Publication Issue :
May-June-2021

Article History
Accepted : 10 May 2021
Published: 16 May 2021

ABSTRACT

Electronic Mail (E-mail) has established a significant place in information user's life. Mails are used as a major and important mode of information sharing because emails are faster and effective way of communication. Email plays its important role of communication in both personal and professional aspects of one's life. The rapid increase in the number of account holders from last few decades and the increase in the volume of mails have generated various serious issues too. The content base mail classification can be classified into four ways namely Private, Public, Newsletter, and Anonymous. Every user has the right to choose their keyword (a semi-private password). Those contacts who know the user's keyword will be classified as private contacts and those users who are unknown them classified anonymous contacts. A contact can be classified as public or private, upon verification of an anonymous contact. Any newsletter or group mails are classified into newsletter contacts. It is highly likely that the rests are junk mail or spam. In this project, a spam detector to identify an email as either spam or ham is built using n-gram analysis. The system involves the classification of mails based on user's contacts. This way any mail from a contact whom the user knows very well is being displayed.

Keywords : Email classifications, Content Based Filtering, Spam detector, Data mining Classification.

I. INTRODUCTION

Internet and email have changed the world with annual growth increasing with the development of mobile technology and the email itself. With the increase in number of Internet users, email is becoming the most extensively used communication mechanism. In recent years, the increased use of emails has led to the emergence and further escalation of problems caused by spam and phishing emails. With the increase in number of Internet users, email is becoming the most extensively used communication mechanism. In recent years, the increased use of emails has led to the emergence and further escalation of problems caused by spam and phishing emails. The email classification process is divided into three distinct levels: pre-processing, learning, and classification.

Though there are several email spam filtering methods in existence, the state-of-the-art approaches are discussed various paper. We used below the

categories of spam filtering techniques that have been widely applied to overcome the problem of email spam.

Content Based Filtering Technique: Content based filtering is usually used to create automatic filtering rules and to classify emails using machine learning approaches, such as Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbor, Neural Networks, N-Gram based feature selection. This method normally analyses words, the occurrence, and distributions of words and phrases in the content of emails and used then use generated rules to filter the incoming email spam.

II. LITERATURE REVIEW

[1] Bo Yu and Zong-ben Xu (2008) performed a comparative analysis on content-based spam classification using four different machine learning algorithms. This paper classified spam emails using four different machine learning algorithms viz. Naıve Bayesian, Neural Network, Support Vector Machine and Relevance Vector Machine. The analysis was performed on different training dataset and feature selection. Analysis results demonstrated that NN algorithm is no good enough algorithm to be used as a tool for spam rejection. SVM and RVM machine learning algorithms are better algorithms than NB classifier. Instead of slow learning, RVM is still better algorithm than SVM for spam classification with less execution time and less relevance vectors.

[2] Loredana Firte, Camelia Lemnaru and Rodica Potolea (2010) performed a comparative analysis on spam detection filter using KNN Algorithm and Resampling approach. This paper make use of K-NN algorithm for classification of spam emails on predefined dataset using feature's selected from the content and emails properties. Resampling of the datasets to appropriate set and positive distribution was carried out to make the algorithm efficient for feature selection .

[3] Megha Rathi and Vikas Pareek(2013) performed an analysis on spam email detection through Data Mining by performing analysis on classifiers by selecting and without selecting the features.

[4] Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta and Anuja Arora (2014) performed a comparative analysis on text and images by using KNN, Naïve Bayes and Reverse-DBSCAN Algorithm for email spam detection. This analysis paper proposed a methodology for detecting text and spam emails. They used Naïve Bayes, K-NN and a modified Reverse DBSCAN (Density- Based Spatial Clustering of Application with Noise) algorithm's. Authors used Enron dataset for text and image spam classification. They used Google's open source library, Tasseract for extracting words from images. Results show that these three machine learning algorithms gives better results without preprocessing among which Naïve Bayes algorithm is highly accurate than other algorithms.

[5] Savita Pundalik Teli and Santosh Kumar Biradar (2014) performed an analysis on effective email classification for spam and non-spam emails.

[6] Ali Shafigh Aski and Navid Khalilzadeh Sourati (2016) performed an analysis using Machine Learning". This paper utilized three machine learning algorithms viz. Multi-Layer Neural Network, J48 and Naïve Bayes Classifier for detection of spam mails from ham mails using 23 rules. The model demonstrated high accuracy in case of MLP with high time for execution while Naïve Bayes showed slightly less accuracy than MLP and also low execution time .

III. PROBLEM DEFINITION

In the recent era, the high use of emails has led to the issues and facing an escalation of problems caused by spam and phishing emails to the user. To avoid this kind of issue we need to classify or sort out mails in a proper manner so that user can easily access their important & less important mails & remove junk mails from the mailbox.

IV. PROPOSED SYSTEM

N-GRAM BASED FEATURE SELECTION

The N-Gram feature selection is a dimensionality reduction method which is used for better classification results by selecting the most desirable feature from the preprocessed data. In this survey N-Gram based feature selection is discussed. N-Gram is a prediction based algorithm used for predicting the chance of occurrence of next word after making observations of N-1 words in a sentence or text corpus. N-Grams uses probability based methods for the prediction of next word. N-Gram is used in text mining and natural language processing. N-grams are the group of co-occurring words that move one or X (number of words in a corpus) steps or words ahead while computing N-Grams. Let X, be the number of words in a given text corpus T, the number of N-Grams can be calculated by: N grams(T) = X – (N-1) (1).

N-Grams are collected from a text corpus and vary according to the size of N. In the above equation for the calculation of N-Gram, T represents the text, X represents the number of words in the text corpus, and N represents the size of the text.S

1) **Uni-Gram**: The N-Gram of size one is termed as Uni-Gram. For example, the word “FOOD” in Uni-Gram can be represented by moving one step ahead viz. “F to O”, “O to O”, “O to D”
2) **Bi-Gram or Di-Gram**: The N-Gram of size two is termed as Di-Gram. For example, “FOOD” in Bi-Gram can be represented by moving two steps ahead in the string of data viz. “FO to OO”, “OO to OD”.
3) **Tri-Gram**: The N-Gram of size three is termed as Tri-Gram and so on for N= 4, N=5 etc. N-Gram for a text corpus “Children are enjoying the sunny weather” using Bi-Gram (N=2) will be “Children are”; “Are enjoying”; “Enjoying the”; “The sunny”; “Sunny weather”

V. FUTURE SCOPE

It is desirable in future that email servers or applications should include different types of pre-defined folders. The first category includes the general traditional folders: Mailbox, sent, trash, etc. It should also allow users to add new folders that can be user defined as well as intelligent or context aware. In convergence with social networks, users should be able to classify emails based on senders or content into different groups.

VI. CONCLUSION

Classification utilizes several natural language processing and data mining activities such as Text parsing, stemming, classification, clustering, etc. There are many goals or reasons why to cluster or classify emails. This may include reasons such as Spam detection, subject or folder classification, etc. Several approaches are evaluated to cluster the emails based on their contents. Classification algorithms are conducted to evaluate the performance of experimented cluster algorithms. The classification of the mail is shown to be the best in the case of N-Gram-based clustering and classification. N-gram feature selection classification technique classify the mail with their gram based unique result. N-Gram is a process that includes dividing the whole text content into sub-terms based on the gram size. Such accuracy can also depend on the number of folders in the classification scheme. It is desirable in the future that email servers or applications should include different types of pre-defined folders. The first category includes the general traditional folders: Mailbox, sent, trash, etc. also allow users to add new folders that can be user-defined. In convergence with social networks, users should be able to classify emails based on senders or content into different groups.

VII. ACKNOWLEDGEMENT

The success of any work depends on efforts of many individuals. We would like to this opportunity to express our deep gratitude to those who extended their support and have guided us to complete this project work.

We wish to express our sincere and deepest gratitude to our guide

Er. Sonal R. Chakole able and unique guidance. We would also like to thank her for the constant source of help, inspiration and encouragement in the successful completion of project. It has been our privilege and pleasure to work under her expert guidance.

We like to thank **Dr.V.P. Balpande (HOD)** for providing us the necessary information about topic. We would again like to thank **Dr.A.M.Shende**, Principal of our college, for providing us the necessary help and facilities we needed.

We express our thanks to all the staff members of **CSE Department** who have directly or indirectly extended their kind co-operation in the completion of our Project Report.

VII. REFERENCES

[1]. Awad M., Foqaha M. Email spam classification using hybrid approach of RBF neural network and particle swarm optimization. Int. J. Netw. Secur. Appl. 2016;8(4) [Google Scholar]

[2]. 2.Bouguila N., Amayri O. A discrete mixture-based kernel for SVMs: application to spam and image categorization. Inf. Process. Manag. 2009;45(6):631–642. [Google Scholar]

[3]. Cao Y., Liao X., Li Y. International Symposium on Neural Networks. Springer Berlin Heidelberg; 2004. An e-mail filtering approach using neural network; pp. 688–694. [Google Scholar]

[4]. Mason S. 2003. New Law Designed to Limit Amount of Spam in E-Mail.http://www.wral.com/technolog [Google Scholar]

[5]. Wang X.L. 2005 International Conference on Machine Learning and Cybernetics (Vol. 9, pp. 5716-5719) IEEE; Aug 2005. Learning to classify email: a survey. [Google Scholar]

[6]. Wang X.L. 2005 International Conference on Machine Learning and Cybernetics (Vol. 9, pp. 5716-5719) IEEE; Aug 2005. Learning to classify email: a survey. [Google Scholar]

[7]. Cormack G.V. Email spam filtering: a systematic review. Found. Trends Inf. Retr. 2008;1(4):335–455. [Google Scholar]

Cite this article as :

Er.Sonal Chakole, Sarita Padole, Apurva Kamble, 4Vandana Wadekar, Ankit Dhande, "Content Based E-Mail Classification", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 8 Issue 3, pp. 141-144, May-June 2021. Available at doi : https://doi.org/10.32628/IJSRSET218323
Journal URL : https://ijsrset.com/IJSRSET218323