EMAIL PRIORITIZATION USING MACHINE LEARNING

Swapnil Choudhari, Narayan Choudhary, Sumit Kaware, Ahmed Shaikh, Sangameshwari Maitri

Dept. of Computer Engineering, Trinity Academy of Engineering, Pune, Maharashtra, India.

Abstract- Personal and business users prefer to use email as one of the crucial sources of communication. The usage and importance of e-mails continuously grow despite the prevalence of alternative means, such as electronic messages, mobile applications, and social networks. Finding out the important mails of the all the emails received on the same day is becoming difficult for many users, as the volume of critical emails continues to grow, the need to automate the management of emails increases for several reasons, such as spam e-mail classification, phishing e-mail classification, and multi-folder categorization, among others. To achieve the objective of study, analysis and comprehensive review to explore the classification as per the importance of emails as users need to look at these mails. The main area of classification is the primary inbox which contains some of the very important mails so to prioritize these mails we are using Natural Language Processing, here we are removing the stop-words present in it and assigning weights to the remaining words. By using frequency count, weight, and access time we can prioritize the mails so that it will be efficient for users to look for the important mails only. The research directions, research challenges, and open issues in the field of e-mail classification are also presented for future researchers.

keyword- Email classification, Natural Language Processing, Machine Learning Techniques.

• INTRODUCTION

Emails are one of the most primary communication methods over the internet in today's digital era. People tend to share the desired contents with other people residing at any point of the internetwork via emails. As the internetwork is spreading more and more with more speed due to advancing technologies, emails are becoming more and more common around the world. Many classification algorithms from machine learning were employed to automatically classify incoming emails into different categories based on the contents of emails. Many classification techniques and classification algorithms such as Neural Network (NN), Support Vector Machine (SVM), and Naïve Bayesian (NB) are currently used in various datasets and showing a good classification result. The techniques such as decision tree (J48), Naive Bayesian classifiers, Neural Networks, Support Vector Machine, etc. has various classification efficiency classify incoming emails into different categories based on the contents of emails. Many classification techniques and classification algorithms such as Neural Network (NN), Support Vector Machine (SVM), and Naïve Bayesian (NB) are currently used in various datasets and showing a good classification result. The techniques such as decision tree (J48), Naive Bayesian classifiers, Neural Networks, Support Vector Machine, etc. have various classification efficiencies.

• LITERATURE SURVEY

In 2008, Taiwo Ayodele Rinat Khusainov David Ndzi proposed the paper "Email Classification and Summarization: A Machine Learning Approach" [5] where they have designed and developed a system that summarizes email messages and also groups emails into activities. Our proposed email summarizer and email classifier will classify the mailbox into spam classification. They have presented an overview of the proposed solutions to extract important words in email messages to provide a better summary than simply running the unprocessed message with a machine learning approach.

In 2013, Sujeet More Dr S A Kulkarni proposed a research paper "Data mining with machine learning applied for email deception" [4] where their work mainly focuses on cognitive (spam) words for classification. This feature is sequential unique and closed patterns which are extracted from the message content. They have shown the method to distinguish the spam words and normal words present in the mails. Their method which can be easily implemented, compares amiably with respect to popular algorithms, like Logistic Regression, Neural

Network, Naïve Bayes and Random Forest using polynomial kernel as filter. They have only worked on different spam message classification. They are considering the spam words to be extracted from spam mail and get the linguistic analysis for that mail so that it can be considered as spam mail.

In 2016, Wei Hu, Jinglong Du, and Yongkang Xing proposed a paper "Spam Filtering by Semantics-based Text Classification" [3] where they have used a novel Chinese spam filtering approach with semantics-based text classification technology was proposed and the related feature terms were selected from the semantic meanings of the text content. Both the extraction of semantic meanings and the selection of feature terms are implemented through attaching annotations on the texts layer by-layer. This filter performed well when experimented on a public Chinese spam corpus. Such that they deal with only spam filtering out of the all mails by the semantics based text classification.

In 2018, Eman M. Bahgat , Sherine Rady, Walaa Gad, Ibrahim F. Moawad proposed a research paper "Efficient email classification approach based on semantic methods"[1] Where 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 (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. 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. This approach gives us the idea of using different algorithms together for increasing the efficiency of work that it needs to do.

• PROPOSED SYSTEM

The current method of email prioritization includes categorization based on type, this will only make the user comfortable while looking for email in a specific category. Even if we get prioritized classification of mail, this classification includes the mail from starred mail, mail interaction between person to person. So we get the classified folder in the email box which contains all these mails together named as 'important'. It does not give us the priority of mail according to the content present in that mail. Thus, to make the user aware of the prior emails in between hundreds of unread emails every day. We are using a machine learning approach to define the priority of mail through the content present in it. It will ease checking only the required mail effectively by its priority. Our proposed system will make consumers less time to check for certain important emails. It will be beneficial to all kinds of users whether they are students, businessmen or organizations.

1) All E-mails are directly fetched from a Gmail account.

2) Then all mails will go under pre-processing, in this step information is extracted from an email that contains a subject, from, date, mail body.

3) Apply Machine Learning technique for further process. It includes NLP for identifying important words from mails and their respective meaning after that we eliminate the stop words present in each mail and assign weights to every other word present in that mail.

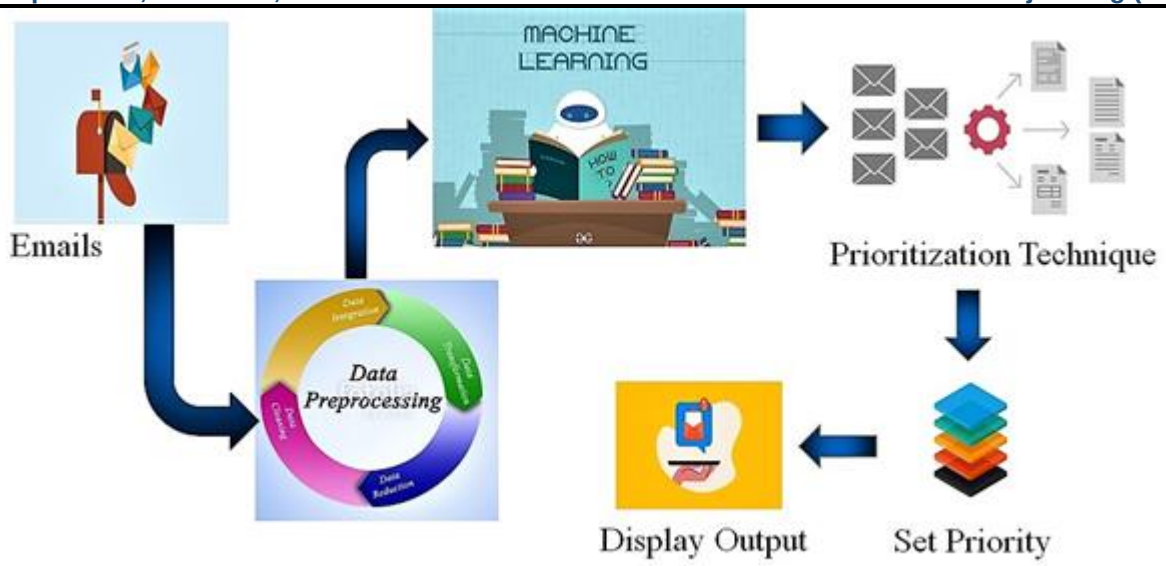

Fig. 1 Proposed System Flow

4) The priority of mails is decided according to frequency count, access time, weighting of words present in that mail through Rank Calculation function.

5) The priority is set by comparing the threshold value of rank and median rank value of mails and such that emails are arranged in a prior order.

• IMPLEMENTATION

The emails are fetched from the Gmail account of one of our team-mate. For fetching the emails we need to first establish the connection and for this, we need to do the authentication of the Gmail account. Refer this site to authenticate the Gmail account. https://developers.google.com/gmail/api/quickstart/python

After that, fetch primary mailbox emails and then the pre-processing starts. The email contains various other things like scripting language, coding of an image, etc. The next Process is to remove this unwanted data that is present in the email. After removing the data goes under different techniques to meet our criteria. The calculation of rank is decided by four parts

1) Frequency count of the email.

2) Access time of the email.

3) Subject of the email.

4) Content of the email

A) The frequency count is calculated by grouping the emails on the sender's email id i.e. Group by ("From"). This will collect all the emails of the sender and give the count of the email that had sent or the total conversation of that sender.

B) Access time of email is the time difference between the users who got the email and the user who read the email.

C) The subject of the email is very important as we can get 80% to 90% information of the email by just reading the subject of the email. The importance of the email is calculated by Natural Language Processing.

D) Like subject line content of the email is also important and hence we also calculated the importance of Natural Language Processing.

• RESULT

The data went through various operations and then we got the output as a prioritized email. For prioritizing the email we also assign the rank to each email. The emails then can be arranged in ascending order to identify the important emails. The result of the data is divided into two parts training data and testing data.

Training Output- For Training data, we had taken 80% part of actual data. The data went through the implementation process. After that, it generates the desired output and this implementation process is again done for testing data. The graphs below are generated which shows frequency (no of mails) against no of users (i.e. no of emails per user).

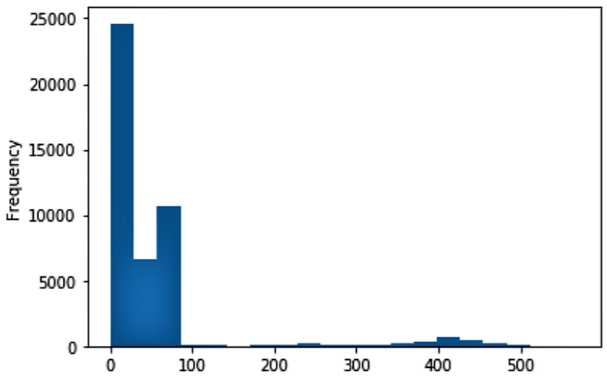

Fig. 2 Email per user vs frequency of emails

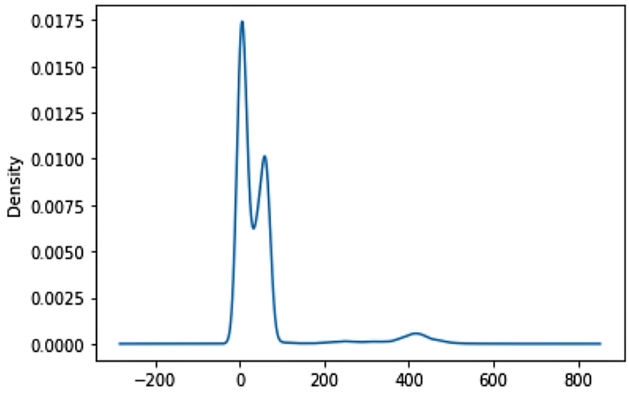

Fig. 3 No. of emails per rank

Testing Output- After creating a model from trained data we apply that model to test data which is 0.2 % of actual data. And finally, we got the rank of each mail from test data with their priority value (i.e. true or false). After Predicting Priority, we have generated graphs for analyzing results. The below graph shows frequency (no of mails) against no of users (i.e. no of emails per user).

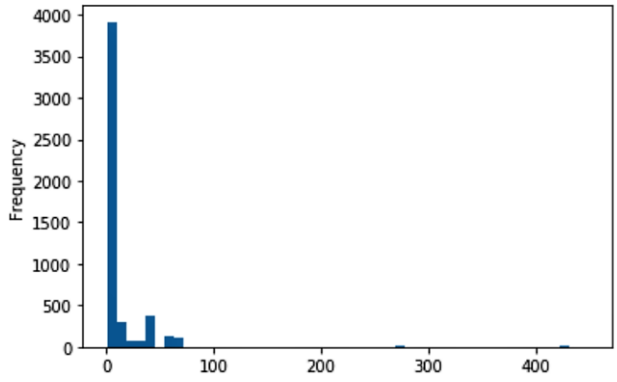

Fig. 4 Email per user vs frequency of emails

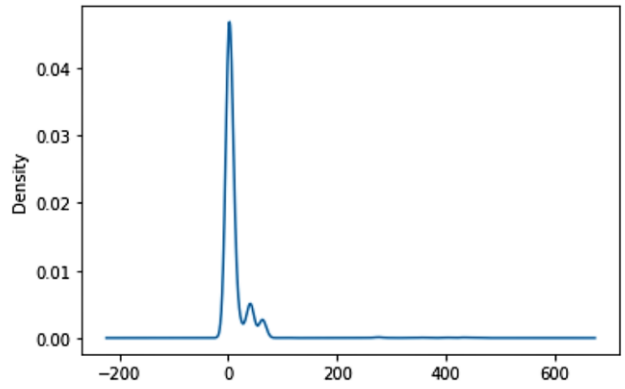

Fig. 5 No. of emails per rank

• CONCLUSION

Thus we have developed the email prioritization system that we are going to use, that will help us to get the important emails for the user so that he/she won't miss out on any information about the important things in their mails. This will also help to make their day to day work better and easy. Priority to each and every mail will give us idea about importance of each mail. We can now check the important mails very efficiently rather than wasting our valuable time on opening and reading the unimportant mails of our inbox

• FUTURE SCOPE

In our next study, we plan to examine how to send notifications for the highest prior email on a user's phone with a short description of the message context which can be very efficient for the user while checking only the prior emails. Attachments present in the email are very complex to read while extraction of information in the pre-processing stage. While extraction of information from email data attachments remains in coded format so it is complex to extract and read the attachments so we are examining the other methods to read and extract information from attachments. Including these features in our system will make this system very useful to any user while accessing the prior emails easily.

• REFERENCES

[1] Eman M.Bahgat, Sherine Rady, Walaa Gad, Ibrahim F.Moawad, "Efficient email classification approach based on semantic methods", Ain Shams Engineering Journal, Volume 9, Issue 4, December 2018, Pages 3259-3269.

[2] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Nahdia Majeed, Mohammed Ali Al-Garadi, "Email Classification Research Trends: Review and Open Issues", Kuala Lumpur 50603, Malaysia, 2017.

[3] W. Hu, J. Du and Y. Xing, "Spam filtering by semantics-based text classification," 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI), Chiang Mai, 2016, pp. 89-94.

[4] More S, Kulkarni S. Data mining with machine learning applied for email deception. International conference on optical imaging sensor and security. IEEE; 2013.

[5] C. Li, W. Song and Y. Yi, " Email Classification Using Semantic Feature Space," in Advanced Language Processing and Web Information Technology, International Conference on, null, 2008 pp. 32-37. doi: 10.1109/ALPIT.2008.93 keywords: {email classification; neural networks; semantic feature space}

[6] W. Li, W. Meng, Z. Tan and Y. Xiang, & quot ;Towards Designing an Email Classification System Using Multi-view Based Semi-supervised Learning,& quot; in 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Beijing, China, 2014 pp. 174-181. doi: 10.1109/TrustCom.2014.26 keywords: {electronic mail; semi supervised learning ;training; supervised learning; data models ; support vector machines ; feature extraction}

[7] I. Moon, Y. Yang, F. Lin and S. Yoo, " Personalized Email Prioritization Based on Content and Social Network Analysis " in IEEE Intelligent Systems, vol. 25, no. 04, pp. 12-18, 2010. doi: 10.1109/MIS.2010.56 keywords: {electronic mail; personalization; clustering; classification; And association rules; feature extraction or construction; Mining methods and algorithms; intelligent systems}

[8] Kobayashi, V. B. et al. (2018) 'Text Classification for Organizational Researchers: A Tutorial', Organizational Research Methods, 21(3), pp. 766–799. doi: 10.1177/1094428117719322.

[9] T. A. Almeida, A. Yamakami, and J. Almeida, "Filtering Spams using the Minimum Description Length Principle," Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1854–1858, 2010.

[10] C. Kreibich, et al. "Spam craft: An inside Look at Spam Campaign Orchestration," Proceedings of the 2nd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more, pp. 4–4, 2009.

[11] G. Forman, and E. Kirshenbaum, "Extremely fast text feature extraction for classification and indexing," Proc of 17th ACM Conference on Information & Knowledge Management, 2008, pp. 1221–1230. J. M. G. Hidalgo, "Evaluating cost-sensitive unsolicited bulk email categorization," ACM Symposium on Applied Computing, 2002, pp. 615–620.