GE-103
Mail Spam Detection

Kirtdeep Kaur, Ayush Singh, Dinesh Swami, Parau Majhi

2021MEB1291, 2021meb1291@iitrpr.ac.in

2021MEB1275, 2021meb1275@iitrpr.ac.in

2021MEB1282, 2021meb1282@iitrpr.ac.in

2021EEB1191, 2021eeb1191@iitrpr.ac.in

Abstract— **Email is one of the most secure mediums of online communication and transfer of data and messages. It is used in many fields of work. We know mails have two categories, namely Ham and Spam. Spam mails are junk or unwanted mails, these types of mail can harm the receiver, and can steal his/her personal information. There is a rapid increase in the number of spam mails day by day. Email spam is an operation to send messages in bulk by mail. So Spam Detection is very much needed to prevent spam. Several machine learning and deep learning techniques are used for this.**

***Keywords*— file handling, functions, loops, string matching(DNA method), string functions**

I. Introduction

As we know email is a primary source of communication. With the passage of time, there is a rapid increase in Spam emails. Spam emails include financial services, adult content, online degree, work-from-home jobs, online gambling, etc. As the name of our project suggests, it helps us to check whether the email sent to you is spam mail or not.

II. Literature Review

In Ref. [1], the technique used was how frequently the words are repeated. Their detection was based on part of speech tagging. The key sentences, those with keywords, of the incoming emails have to be tagged and thereafter the grammatical roles of the entire words in the sentence need to be determined and then they will be put together in a vector to take the similarity between received mails. K-mean algorithm is used to classify the received mails. Vector determination is used to determine the category of the email.

In Ref. [2], the authors used the method that investigates the use of string matching algorithms for spam detection. This work examines and compares the efficiency of six well-known matching algorithms, Longest Common Subsequence (LCS), Levenshtein Distance(LD), Jaro, Jaro Winkler, Bi-gram, and TFIDF on two datasets which are Enron corpus and CSDMC2010 spam database. They also observed that the Bi-gram performs best in spam detection in both datasets.

In Ref. [3], they used machine learning techniques to detect the pattern of the repetitive keywords which are classified as spam. This system also classified the emails based on the parameters contained in their structure as Cc/Bcc, domain, and header. These parameters would be considered as a feature when applying them to the machine learning algorithm. This model is a pre-trained model.

III. Objective

Why is it necessary to combat spam? Spam started a number of years ago with modest advertising mailings, which over time, have developed into a serious technical, economic, and even social threat.

A. *Communications overload*

Spam blocks communication channels and creates traffic that has to be paid for by either the provider or the user (or the employer in the case of a company). Additionally, there are mail servers that receive and process spam, and these servers have to be maintained by highly-paid specialists. Therefore there are substantial infrastructure running costs also.

B. *Waste of time*

If spam reaches a user's inbox, a recipient has to delete it manually. A person who reads 10-20 emails per day may receive in the region of 160-180 spam messages along with their business correspondence. That means that they will spend 5-6 hours per month just deleting spam, to the detriment of their productive working time.

C. *Irritation and discontent*

By having to manually delete spam, a user becomes an ‘electronic’ waste disposal technician. Being forced to take such measures cannot help but irritate the user, resulting in unwanted negative emotions.The loss of an important email that accidentally gets deleted along with the plethora of spam. Everyone who has faced such a situation at any time will immediately understand. No further comment is required.

D. *Criminalization of spam*

Spammers have been most inventive in creating ever more attractive ‘bait’ for the user and seeking new targets for their attacks Hot topics in the news can be used in spam messages to get your attention. In 2020 when the world was facing the Covid-19 pandemic and there was an increase in work-from-home jobs, some scammers sent spam messages promising remote jobs that paid in Bitcoin. During the same year, another popular spam topic was related to offering financial relief for small businesses, but the scammers ultimately asked for bank account details. News headlines can be catchy, but beware of them in regards to potential spam messages.

1. In addition, the services of the spammer are in constant demand by virus writers. Virus writers use spam mailings to distribute their latest creations, often placing links to infected sites within the mailing that are designed to lure the unwary user to click on them for one reason or another. A recipient of such spam thus runs the risk of their computer being infected by a malicious program.
2. According to the experts, the annual overall loss resulting from spam is estimated to be tens of billions of dollars. As a result, anti-spam protection is not only desirable but an urgent necessity. If spammer activity is not restricted, the email could easily become a thing of the past, eclipsed by the overwhelming volume of spam. In the modern world, anti-spam protection and antimalware protection have become an indispensable part of any IT security system.
3. Email spam filters catch many of these types of messages, and phone carriers often warn you of a “spam risk” from unknown callers. Whether via email, text, phone, or social media, some spam messages do get through, and you want to be able to recognize them and avoid these threats.

Therefore our team developed a mail spam checker which is a python program. It compares the messages that have been sent to us with the different type of sentences listed in the excel file and identify whether the message is “SPAM” or not.

IV. Conclusions

This section concludes your project with the results achieved. You can add different test cases and boundary conditions used in testing the project.

Acknowledgment

All the members of the python team” PREP ALPHA” would like to express our special thanks of gratitude to our Introduction to computing and Data Structure professor, Dr. Sudarshan Iyengar gave us the golden opportunity to do this wonderful project on the topic “MAIL SPAM CHECKER”. This topic helped us in doing a lot of research and came to know about many new things, we are really thankful to them.

We would also like to extend our gratitude to all the teaching assistants including our assigned teaching assistant Mr.Raushan Sir who helped us find a new solution to our problem at the last moment. My thanks and appreciation also go to all of my teammates who did really hard work in understanding the concept, gathering sufficient information on this topic, and writing the genuine code to identify the spam messages. We also addressed several difficulties in coordinating the activities of the project but with the support of all of my team members and the constant supervision of Mr.Sudarshan Sir and Mr. Raushan Sir, we completed the project.

Finally, Words are not sufficient to express gratitude to our cherished parents and friends for their cooperation and encouragement as they helped us a lot in finalizing this project within the limited time frame.

References

[1] https://www.geeksforgeeks.org/applications-of-string-matching-algorithms/

[2] https://www.geeksforgeeks.org/file-handling-python/

[3] https://www.programiz.com/python-programming/methods/string

[4]