CSCE 5290: Natural Language Processing
Project Proposal – Spam Filtering

Project Description:

Spam is typically defined as undesired text that is sent or received over social media platforms like Facebook, Twitter, YouTube, e-mail, etc. It is produced by spammers in an effort to divert consumer's attention from social media marketing and malware distribution, among other purposes. Spam filtering is a mechanism used to filter spam messages or emails to prevent its delivery. The volume of unsolicited email has significantly increased, necessitating the development of more robust and efficient antispam filters. Machine learning algorithms have been used to successfully identify and filter spam emails.
Here, along with the control mechanisms and datasets utilized on spam detection, we also discuss the difficulties in identifying spam.

Github repository link: https://github.com/ShishiraRudrabhatla/spam-filtering-nlp

1. Project Title: Spam Filtering using NLP

Team Members: Group # 13

No.	Name	Student Id
1	Shishira Rudrabhatla	11560633
2	Nathisha Marru	11560642
3	Abhishek Rangineni	11435098
4	Venkata Sai Rahul Kumar Katta	11505841

2. Goals and Objectives:

Motivation:

One of the quickest and most affordable modes of communication is now email.
Spam emails, on the other hand, have dramatically increased over the past several years as a result of the rise in email subscribers. Since spammers are constantly looking for ways to get around existing filters, new filters must be created to stop spam. It is due to this major increase in spam emails that has led to the necessity of this project with an aim to train spam-filtering models better and stay up to date with the latest tricks spammers are coming up with every day and try to filter emails in the most effective ways possible.

Significance

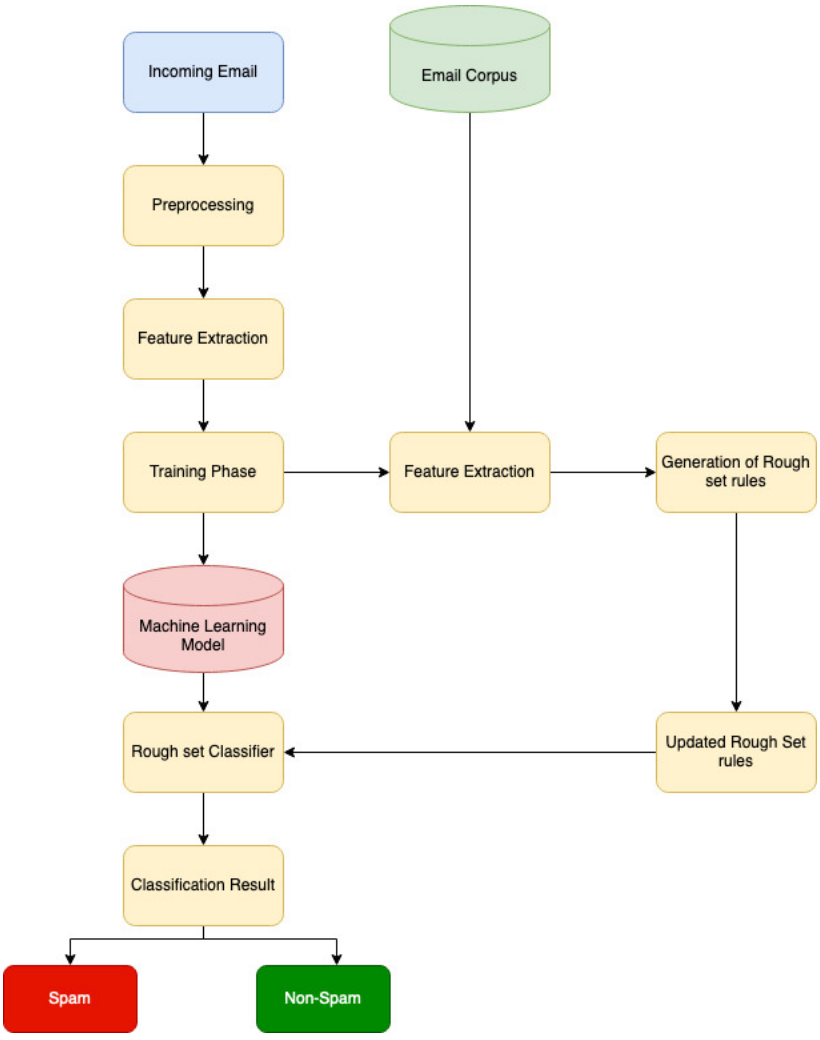
It is critical to take extra precautions to safeguard your equipment, especially if it processes sensitive information like user data. This is since clicking on a spam email might put your

computer and personal information at risk of being infected with malware. Spam filtering implementation is crucial for every firm as it does not only helps keep junk out of email inboxes, but it also improves the functionality and usefulness of business communications by ensuring that they are only utilized for what they were intended for. Since many email-based attacks attempt to deceive users into clicking on a malicious file, asking them for their credentials, and other information, spam filtering becomes one of the most needed functionalities in such cases.

Objectives:

Typically, text classification is the primary methodology used for email screening. There are several alternatives available when determining how to include machine learning into a Python email categorization. The objective is to develop a Python-based spam filtering approach in which relevant spam are first filtered from the training dataset and then used to generate training and testing tables for various data mining algorithms.

Workflow diagram: Idea of workflow diagram



Features:

The proposed spam filtering system is expected to have the following features once fully implemented:

- Effective performance supported by the advantages and powers in natural language processing.
- Ability to categorize your emails as spam by training the classifier on your own dataset.
- Simplicity in usage.
- Fast performance once the training of the classifier is completed.
- Better model with higher accuracy in filtering spam e-mails.

3. References:

https://www.kaggle.com/datasets/nitishabharathi/email-spam-dataset?select=lingSpam.csv

https://www.kaggle.com/datasets/karthickveerakumar/spam-filter

https://www.sciencedirect.com/science/article/pii/S2405844018353404#fig5

Debnath, K., & Kar, N. (2022, May). Email Spam Detection using Deep Learning Approach. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)* (Vol. 1, pp. 37-41). IEEE.

Ismail, S. S., Mansour, R. F., El-Aziz, A., Rasha, M., & Taloba, A. I. (2022). Efficient E-Mail Spam Detection Strategy Using Genetic Decision Tree Processing with NLP Features. *Computational Intelligence and Neuroscience*, *2022*.