**Project Name:** Spam EMail Classification

**Github Link:** https://github.com/projectsforstudents2022/Spam\_EMail\_Classification.git

**Why was this project created?**

Spamming emails is currently one of the main problems that everyone in the online community must deal with. Because email is a simple means of communication and is inexpensive, people in this society like to use it to share files and information. However, by sending spam emails, these emails primarily harm both individuals and professionals. The frequency of spam emails and texts is rising daily. These spam emails are typically sent by individuals looking to market themselves or make money. For people who get the spam mail, the rising volume of it clogs up traffic and wastes their time.

**What problem is it solving?**

The project's primary goals are to detect spam emails and to improve data storage. In this suggested system, the two filtering models are used to detect spam emails. One is Opinion Rank, which employs the high page rank and inverse page rank algorithms and is dependent on how trustworthy the mail id is. The Opinion Rank will function by averaging these findings and computing their mean.

**Entire explanation of project**

* **PROPOSED APPROACH**

Spam Mail Detection System for sorting emails into spam and ham messages. The hybrid bagged method concept was used to introduce the system's powerful categorization capabilities. The correlation-based feature selection technique and the unique hybrid bagging strategy are used for the feature selection and classification, respectively. The Spam mail detection system loads an email dataset. From the Ling spam dataset, several emails are chosen at random. For the purpose of categorization, the dataset consists of a total of 1000 emails, both spam and ham emails.

Three phases make up the pre-processing stage. To begin with, the text data is tokenized. The statement is broken up into token words. Stop words are eliminated from the tokenized words. Stop words are undesirable words without any linguistic significance. Approximately 670 stop words are manually added to a text file, and during pre-processing, words are taken out of the text. The stemming process is the third stage in the pre-processing module. The term is reduced to its basic word by the stemming process. Text data with a feature set is initially thought of as a collection of words. The term frequency approach is used to estimate the word count in a text. All words are given a frequency value, and those with frequencies below a certain threshold are disregarded. This technique both shows how useful a term is and minimizes the search space.

Algorithm for creating next word prediction model :

**Step 1:** Import Libraries & Load Dataset

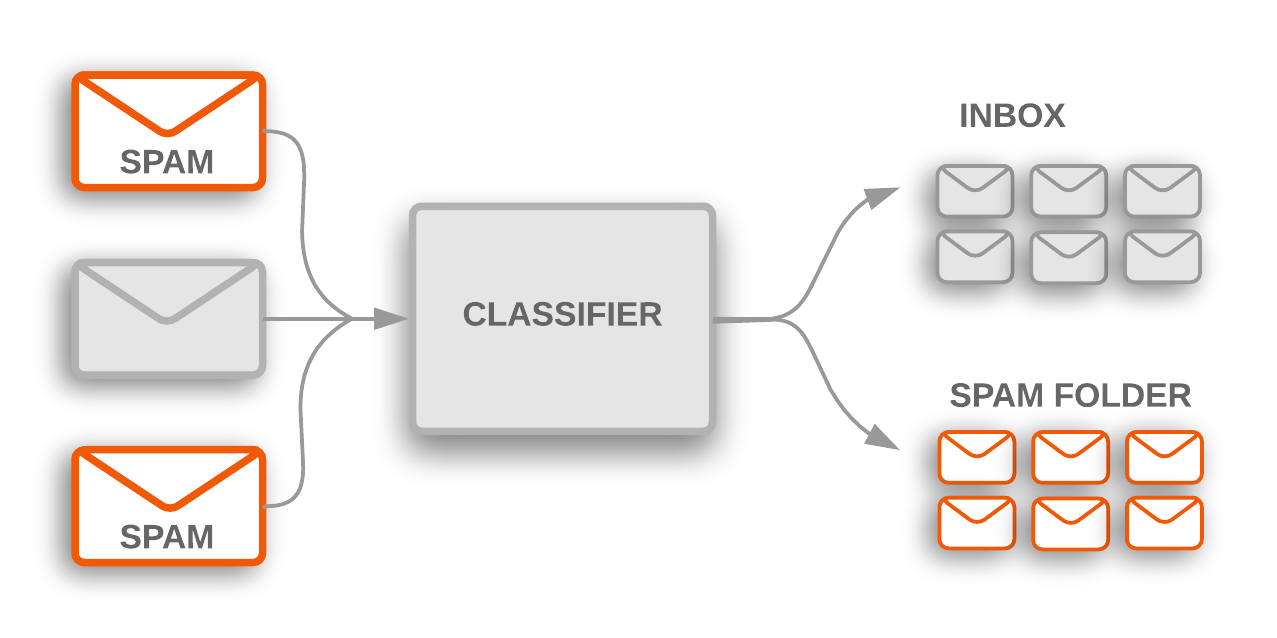
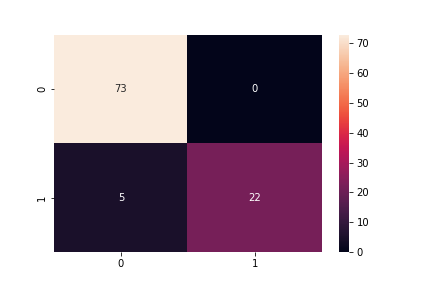
**Step 2:** Stop Words Handling

**Step 3:** Tokenization

**Step 4:** Build SklearnClassifier Model

**Step 5:** Train Model

**Step 6:** Testing & Visualization

* **DATA FLOW DIAGRAM**
* **RESULT**
* **CONCLUSION**

This system's suggested method classifies emails as "Spam" or "Not Spam" and also optimizes data after receiving emails as input. This is the most well-liked Topic Modeling Approach and a generative probabilistic model. The emails that were utilized as input are categorized using this into several pertinent subjects. The OpinionRank algorithm gathers URLs from emails and uses them as input to determine if a website is trustworthy.