CSCE 5290.003 (13149)

Natural Language Processing

Spam Email Detection with Text Classification and Sentiment Analysis

Using Machine Learning Random Forest

Project Proposal

SPAM Email Detection with Sentiment Analysis

Project Proposal

1. Project Title and Team Members

Project Group **- 5**

Project Title **- Spam Email Detection with Sentiment Analysis**

Team Members:

- **Danda Reethika Reddy - 11608030**
- **Surya Sai Raj Lakkoju - 11610081**
- **Ashesh Pinisetti - 11649062**
- **Mohammad Khaja Moinuddin - 11603687**

2. Goals and Objectives:

- **Motivation:**
  The reason behind getting motivated about this project is that in the current existing machine learning models we see there are models that can detect whether the emails are spam or not, but not about categorizing which type of email that content actually defines. Here we analyze how the spam emails were detected using the Text classification techniques and then we will be implementing the sentimental analysis techniques to find out what kind of email it is for example: Threaten, Marketing, Phishing or etc. Here we will be using a random forest Machine learning model which will be trained with the dataset to analyze the accuracy.

- **Significance:**
  The main significance of our project is to identify and classify the spam emails. Identifying the spam emails will help users in not getting involved in any type of fraud. The users will have a basic understanding of what the email is about if it is categorized as marketing, anti-virus, phishing, or money scam spam.

  Importance for spam mails to be filtered is, as it can contain malicious content that can spread viruses and cyber attacks.To overcome these types of attacks, spam mails filtering is necessary.

After identification of spam emails,Based on the content present in the mail we are categorizing them into groups by using text recognition and sentiment analysis.

- **Objective:**
  The primary objective of spam email detection with sentiment analysis is to identify and filter out unwanted emails from the inbox of an email account. This involves using machine learning algorithms to automatically detect and categorize emails as spam or not.

  Sentiment analysis, on the other hand, involves analyzing the language used in a piece of text to determine the underlying emotional tone or sentiment. In the context of email filtering, sentiment analysis can be used to determine whether an email is trying to manipulate or deceive the recipient, such as through the use of overly positive or negative language.

  By combining spam email detection with sentiment analysis, it is possible to create a more effective email filtering system that can identify and block unwanted emails that may be attempting to manipulate or deceive the recipient. This can help improve the user's email experience by reducing the amount of unwanted or malicious emails that they receive, and also enhance their security and privacy online

- **Features:**
  Here are some common features we are using in spam email detection:

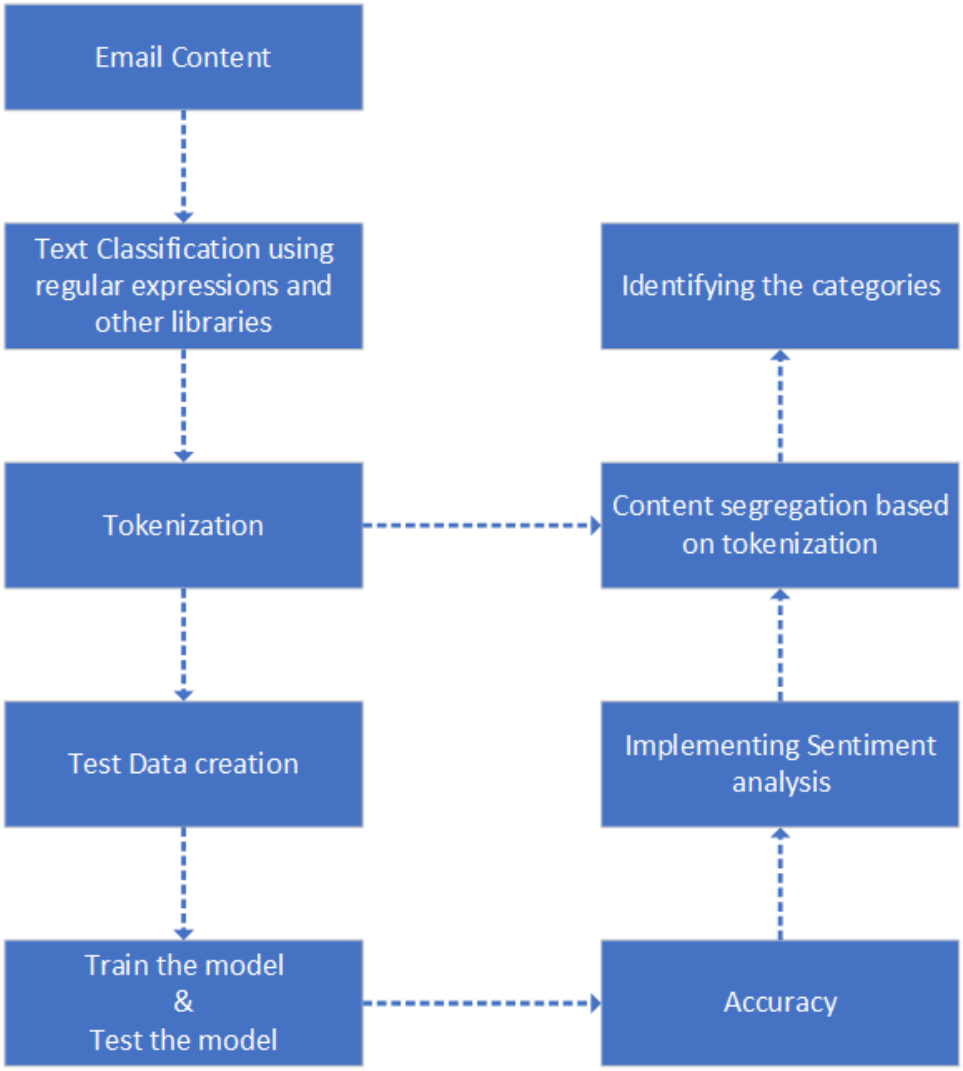
  Content Analysis: Spam emails often contain certain patterns or phrases that can be identified and used as indicators of spam. These include things like suspicious keywords, unusual formatting, and overuse of capital letters or exclamation points.

  Blacklists and Whitelists: These are lists of known good and bad email addresses, domains, and IP addresses. Emails from whitelisted sources are allowed to pass through, while those from blacklisted sources are blocked or flagged as spam.

  Reputation Analysis: The reputation of an email sender can be evaluated based on factors such as how often their emails are marked as spam, the types of content they send, and the quality of their email infrastructure.

  URL Analysis: Links in spam emails often lead to fraudulent or malicious websites. URL analysis can be used to identify these links and block or flag emails that contain them.

Architectural flow diagram:



3. References:

- https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning?select=Phishing_Legitimate_full.csv
- https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
- https://academic.oup.com/jigpal/article/28/1/83/5680435
- https://link.springer.com/book/10.1007/978-3-642-25237-2#page=202
- https://www.mililink.com/upload/article/2008948707aams_vol_215_march_2022_a27_p2695-2705_jasneet_kaur.pdf

4. Git Repository URL: https://github.com/surya5a72/csce5290_NLP