EMAIL SPAM DETECTION USING MACHINE LEARNING ALGORITHMS

Overview

The Email Spam Detection project aims to build a machine learning model capable of distinguishing between legitimate ("ham") and spam emails. The project involves several key steps including data preprocessing, exploratory data analysis (EDA), text processing, model building, and evaluation.

The aim of this developed system which can perform early detection of spam mails with a higher accuracy by combining the results of different machine learning techniques. The algorithms like Neural Networks ,Naive Bayes ,Support vector classifier ,Logistic Reggression are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for Email Spam Detection.

Keywords:

• Neural Networks

• Naive Bayes

• Support vector classifier

• Logistic Regression

• Spam email

1. Data Processing

The dataset used for this project consists of labelled emails, categorizing them as either spam or ham. Initial data processing steps include loading the dataset and gaining an understanding of its structure.

- Dataset Shape: (5572, 2)

2. Exploratory Data Analysis (EDA)

EDA involves gaining insights into the dataset to better understand its characteristics and distributions. Key EDA steps include:

- Data Visualization: [Include visualizations to show distributions, e.g., pie chart of spam vs. ham]
- Statistical Analysis: [Include any relevant statistics or insights gained from the EDA]

3. Text Processing

Text processing is a crucial step in preparing the data for model building. This involves a series of operations including:

- Lowercasing: Convert all text to lowercase for uniformity.
- Tokenization: Splitting text into individual words or tokens.
- Removing Special Characters: Eliminating non-alphanumeric characters.
- Removing Stop Words: Removing common words that do not carry significant meaning.
- Stemming: Reducing words to their root form.

A)Tokenization:

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called Tokens. The list of Tokens becomes input for further processing such as parsing or text mining.

B)Removal of Stop Word:

Sometimes, the extremely common word which would appear to be of very little value in helping select documents matching user need are excluded from the vocabulary entirely

C) Stemming and Lemmatization:-

Stemming is the process of removing the last few characters of a given word,to obtain a shorter form, even if that form doesn't have any meaning. Lemmatization in linguistics,is the process of grouping together the different inflected forms of a word so they can be analysed as a single item.

4. Feature Selection and Extraction:

- Number of Capitalized words.
- Sum of all the character length of words.
- Number of words containing letters and numbers.
- Max of ratio of digital characters to all the characters of each word.
- Sum of all the character lengths of words, etc...

5. Model Building

In this project, multiple machine learning algorithms were employed for the task of email classification. The models used include:

Naive Bayes

Gaussian Naive Bayes

- Description: Gaussian Naive Bayes is suitable for classification tasks with continuous features.
- Performance:
  - Accuracy score: 0.8556053811659193
  - Confusion matrix:

    [[822 135]

    [26 132]]

  - Precision score:0.4943820224719101

Multinomial Naive Bayes

- Description: Multinomial Naive Bayes is designed for discrete features, making it suitable for text classification.
- Performance:
  - Accuracy score: 0.9748878923766816
  - Confusion matrix:

    [[947 10]

    [18 140]]

  - Precision score: 0.9333333333333333

Bernoulli Naive Bayes

- Description: Bernoulli Naive Bayes is used for binary feature classification tasks.
- Performance:
  - Accuracy score: 0.9623318385650225
  - Confusion matrix:

    [[952   5]

    [37 121]]

  - Precision score: 0.9603174603174603

Logistic Regression

- Description: Logistic Regression is a linear classification algorithm that models the probability of a binary outcome.
- Performance:
  - Accuracy score: 0.9650224215246637
  - Confusion matrix:
    [[957 0]
    [39 119]]
  - Precision score: 1.0

Support Vector Machine (SVM)

- Description: SVM is a powerful classification algorithm that finds an optimal hyperplane to separate classes.
- Performance:
  - Accuracy score: 0.9659192825112107
  - Confusion matrix:
    [[949 8]
    [30 128]]
  - Precision score: 0.9411764705882353

Each model was trained on the preprocessed data and evaluated for its performance.

6. Model Evaluation

Model performance was assessed using various metrics including accuracy, confusion matrix, and precision score. The evaluation results are as follows:

Result Analysis:

	Accuracy score	***Precision score***
Support vector machine	0.96	0.94
Navie bayes	0.97	0.96
Logistic regression	0.96	1.0

7. Building a Predictive System

A predictive system was developed to demonstrate the model's practical application. Users can input a message, and the system will classify it as spam or ham.

Future Improvements

Potential enhancements for this project include:

- Experimenting with different text processing techniques.
- Exploring more advanced machine learning algorithms.
- Fine-tuning hyperparameters for improved performance.

Conclusion

The Email Spam Detection project successfully achieved its objective of classifying emails as spam or ham using machine learning techniques. The detection of spam at a place close to the sending server is an important issue in the network security and machine learning techniques have a very important role in this topic. In this paper, review of some machine learning techniques used in spam filters and presented challenges faced by these techniques is presented along with an evaluation in terms of various metrics of three machine learning algorithms namely Naïve Bayes, Support Vector Machine, Logistic Regression. The evaluation was based on data set of e-mail messages collected from different e-mail accounts located on different e-mail servers. Although all learning classifiers showed ability to learn but the NB classifier based filters showed better performance in terms of all measuring parameters. However, none of these classification techniques showed 100% predicative accuracy. The dynamic structure of spam and the reaction of spammers towards spam filters makes spam filtering an active area for research and thus there exists a wide scope for development of new spam filters and improvements in the existing ones has potential applications in email filtering and cybersecurity.