Project Report

Spam Mail Prediction Using Logistic Regression

PrernaGati
TECHNOLOGY
ISO 9001:2015 MSME BUSINESS FORUM INDIA

&

Under the Guidance of PGTech

By Kanishk Thakur

Abstract

Spam mail prediction is essential for enhancing email communication efficiency and reducing distractions. This project utilizes logistic regression to develop a model that identifies spam emails accurately. Guided by PGTech institute, I created a robust spam mail prediction system, assessed its performance, and provided recommendation for future improvements.

Introduction

Email is a primary communication tool, but the prevalence of spam emails poses significant challenges. Accurate spam detection can mitigate these issues, enhancing user productivity and security. This project employs logistic regression to classify email as spam or not spam based on their content.

Objectives

1. Data Collection : Obtain a dataset of labelled emails(spam and ham).
2. Data Preprocessing : clean and preprocess the data for analysis.
3. Model Development : Develop a logistic regression model for spam prediction.
4. Evaluation : Evaluate the model’s performance using standard metrics.

Libraries Used

The following libraries were used to implement and evaluate the spam mail prediction model:

1. Pandas: For data manipulation and analysis.

2. NumPy: For numerical computations.

3. Scikit-learn: For machine learning algorithms and model evaluation.

4. Matplotlib/Seaborn: For data visualization.

Methodology

Data Collection

I utilize a csv file named “mail_data.csv”, it contains labelled data of emails. Spam mail is labelled as spam in category and non spam as ham. This dataset is good for logistic regression model.

Data Preprocessing

1. Text Cleaning : Removed HTML tags, special characters, and stop words from the emails.
2. Tokenization : Split the email text into individual words(tokens).
3. Vectorization : Converted text data into numerical data using Term Frequency-Inverse Document Frequency(TF-IDF).

Model Development

1. Feature Selection: Selected relevant features based on their correlation with target variable(spam or ham)
2. Logistic Regression Model : Implemented the logistic regression algorithm using Python’s scikit-learn library.

Evaluation

1. Training and Testing : Split the dataset into training (80%) and testing (20%) sets.
2. Metrics: Used accuracy score to evaluate the model performance.

Results

The logistic regression model achieved and accuracy of 96.5%, with a high accuracy it indicates its effectiveness to distinguish between spam mails and non spam mails. The model is neither underfitting nor overfitting as both training and testing data have almost same accuracy.

Conclusion

This project successfully developed a logistic regression model for spam mail prediction, achieving high accuracy and reliability. Under the guidance of PGTech Institute, I demonstrated the effectiveness of logistic regression for this classification test. Future work will focus on optimizing the model further and exploring advance machine learning techniques.

Acknowledgements

I express my gratitude to PGTech Institute for their guidance and support throughout this project. Special thanks to my mentors and peers who provided valuable insights and feedback.

Kanishk Thakur