Kasetsart University Sriracha Campus

Project Report

Spam Email Detection Using TF-IDF and Logistic Regression

Jirayu Choonade , Witchayut Withichai , Sasitorn Sangpet , Nalintorn Sopapiriyatorn , Saowaluk Praphasirisulee

Bachelor of Science Program in Digital Science and Technology

Abstract

This project presents the development of a spam email detection system using machine learning techniques. The proposed system applies Term Frequency–Inverse Document Frequency (TF-IDF) for text feature extraction and Logistic Regression for binary classification. The dataset used in this study is obtained from Kaggle and consists of labeled email messages categorized as spam or non-spam. The system is designed to handle class imbalance through weighted loss adjustment and decision threshold tuning. Experimental results demonstrate that the proposed approach provides effective performance in spam classification while maintaining computational efficiency.

1. Introduction

Email communication remains a critical medium for personal and professional interaction. However, the increasing volume of spam emails poses significant challenges, including security risks, productivity loss, and privacy concerns. Automated spam detection systems are therefore essential to filter unwanted messages. This project focuses on building a machine learning-based spam email classification system using classical natural language processing techniques and supervised learning.

2. Problem Statement and Objectives

2.1 Problem Statement

Spam email classification is a binary classification problem characterized by highly imbalanced data distributions. Traditional rule-based systems lack adaptability and often fail to generalize to new spam patterns. Machine learning approaches offer a more robust and scalable solution.

2.2 Objectives

The objectives of this project are as follows:

- To develop a spam email detection system using machine learning
- To apply TF-IDF for textual feature extraction
- To train and evaluate a Logistic Regression classifier
- To analyze system performance using standard evaluation metrics

3. Dataset Description

The dataset used in this project is sourced from Kaggle and contains labeled email messages. Each record consists of the email text and a binary label indicating whether the email is spam or not. The dataset is preprocessed to retain only relevant attributes required for text classification.

4. System Architecture

4.1 Overall Architecture

The system architecture consists of five main stages: dataset acquisition, data preprocessing, feature extraction, model training, and model evaluation. The entire system is implemented using Python and widely adopted machine learning libraries.

4.2 System Flow

The system flow begins with loading the dataset, followed by text preprocessing and TF-IDF vectorization. The extracted features are then passed to a Logistic Regression classifier for training. Finally, the trained model is evaluated using validation and test datasets.

5. Data Preprocessing

Data preprocessing includes text selection, label encoding, and dataset splitting. The dataset is divided into training, validation, and test sets using stratified sampling to preserve class distribution. Label encoding is applied to convert categorical labels into numerical form suitable for machine learning algorithms.

6. Feature Extraction Using TF-IDF

TF-IDF is employed to convert textual data into numerical feature vectors. Character-level n-grams ranging from three to six characters are used to capture morphological patterns common in spam messages. Feature dimensionality is controlled through minimum and maximum document frequency thresholds.

7. Machine Learning Model

7.1 Logistic Regression Classifier

Logistic Regression is selected due to its efficiency, interpretability, and strong performance in linear text classification tasks. L2 regularization is applied to prevent overfitting, and class weighting is introduced to address data imbalance.

7.2 Hyperparameter Configuration

Key hyperparameters include the regularization strength, solver selection, maximum iteration limit, and class weight adjustment. These parameters are empirically tuned to achieve optimal validation performance.

8. Model Training

The model is trained using the TF-IDF feature vectors derived from the training dataset. GPU availability is detected for system compatibility, although the training process primarily relies on CPU-based computation. The training process converges after multiple iterations to ensure model stability.

9. Decision Threshold Optimization

Instead of using the default classification threshold, a custom probability threshold is applied to improve spam detection performance. This approach enhances recall for spam emails while maintaining acceptable precision levels.

10. Evaluation Metrics

System performance is evaluated using accuracy, precision, recall, and F1-score. A confusion matrix is also generated to analyze classification behavior across spam and non-spam categories.

11. Experimental Results

Experimental results indicate that the proposed system achieves high classification accuracy and balanced precision-recall performance. The threshold adjustment significantly improves spam detection capability, particularly for minority classes.

12. Model Deployment and Persistence

To support reuse and deployment, the trained Logistic Regression model, TF-IDF vectorizer, label encoder, and classification threshold are serialized and saved using joblib. This enables efficient model loading and inference in real-world applications.

13. Discussion

The results demonstrate that classical machine learning techniques remain effective for spam email detection tasks. Character-level TF-IDF features capture subtle textual patterns commonly found in spam messages. However, performance may degrade when encountering unseen spam strategies.

14. Limitations and Future Work

The current system does not incorporate semantic context or deep learning techniques. Future work may involve integrating word embeddings, transformer-based models, or real-time email filtering systems to improve robustness and adaptability.

15. Conclusion

This project successfully demonstrates the design and implementation of a spam email detection system using TF-IDF and Logistic Regression. The proposed approach achieves reliable performance and provides a solid foundation for further research and development in email security applications.