T-10: Spam Detection Project Report

Group 14
Kanav Singh Chouhan (220495)
Nagisetty Vinay (220678)
Pawan Dhakar (220763)
Sambuddha Chakrabarti (220947)
Tanmay Soni (221130)

***Abstract*—This report provides a comprehensive overview of a spam detection system based on logistic regression. Our objective is to classify text messages as spam or ham (non-spam) accurately. We describe data cleaning and feature extraction techniques, outline the model architecture—including regularization and class weighting strategies—and present performance evaluations based on key metrics. The experimental results demonstrate the model's efficacy and highlight opportunities for further improvement.**

***Index Terms*—Spam Detection, Logistic Regression, Preprocessing, Text Classification, Performance Metrics**

I. Introduction

The pervasive issue of spam in digital communications poses a challenge to both user experience and system security. This project sets out to construct a robust spam detection model capable of distinguishing spam messages from legitimate ones. Given its simplicity and interpretability, logistic regression was chosen as the classifier. The model is trained using a set of text messages that have undergone extensive preprocessing to enhance feature quality. In this report, we delve into each phase of the development process, including data preprocessing, model construction, and performance evaluation.

II. Data and Preprocessing

An essential component of any machine learning project is effective preprocessing. Our dataset comprises thousands of text messages, each labeled as either spam or ham. The following steps ensure that the data is both clean and useful for model training.

A. Data Cleaning

- **Normalization:** All text is converted to lowercase to maintain consistency.
- **Noise Removal:** Special characters, numbers, and extraneous spaces are removed.
- **Text Correction:** Minor typographical errors are addressed where possible.

B. Tokenization and Feature Extraction

After cleaning, each message is tokenized into individual words. Stopwords (e.g., "the", "is", "and") are removed since they offer little discriminative power. The cleaned tokens are then transformed into numerical features using a Bag-of-Words technique implemented via `CountVectorizer`. This vectorization converts text into a sparse matrix representation, which serves as input to the classifier.

C. Data Splitting and Class Imbalance Handling

To evaluate model performance, the complete dataset is segmented into a training set (80%) and a testing set (20%). Additionally, due to the inherent class imbalance—with fewer spam messages than ham—a higher weight is assigned to spam during training. This weighting helps reduce false negatives and ensures that the classifier remains sensitive to the minority class.

III. Model Architecture

The logistic regression classifier forms the backbone of our spam detection system. Key features of the model include:

- **Loss Function:** Binary cross-entropy is used to measure the divergence between predicted and actual labels.
- **Regularization:** L2 regularization is applied to prevent overfitting, ensuring that the model generalizes well to unseen data.
- **Hyperparameters:** A maximum of 2000 iterations has been set to secure convergence. Other hyperparameters remain at their default values.
- **Class Weights:** The class imbalance is mitigated by increasing the penalty for misclassifying spam messages.

The processed bag-of-words vectors pass through the logistic regression model, generating a probability that a given message is spam. A threshold of 0.5 is established for final classification.

IV. Experimental Evaluation

A. Performance Metrics

To assess the classifier's effectiveness, we utilize the following metrics:

- **Accuracy:** The ratio of correctly classified messages to the total number of predictions.
- **Precision:** The proportion of true spam messages among all messages predicted as spam.
- **Recall:** The fraction of actual spam messages that were correctly identified.
- **F1-Score:** The harmonic mean of precision and recall.

This work did not receive any specific grant from funding agencies.

B. Results and Analysis

The logistic regression model achieved strong performance across the evaluated metrics:

- Accuracy: 98.99%
- Precision: 96.61%
- Recall: 95.80%
- F1-Score: 96.20%

In addition, the classifier's confusion matrix is as follows:

$$\begin{bmatrix} 769 & 4 \\ 5 & 114 \end{bmatrix}$$

This matrix indicates that while the model is highly effective overall, there are minor misclassifications which could be the focus of future enhancements.

C. Discussion

The experimental results verify that logistic regression is a viable approach for spam detection when supported by robust preprocessing. The high precision and recall values attest to the model's capability in minimizing false positives and negatives, respectively. However, subtle misclassifications suggest that augmenting our feature extraction process—such as incorporating TF-IDF measures or word embeddings—might further enhance performance. Additionally, exploring alternative classifiers or ensemble methods could provide valuable improvements for more complex or ambiguous cases.

V. Conclusion and Future Work

This report has detailed the construction of a logistic regression model for spam detection, emphasizing the critical roles of data preprocessing and model tuning. The overall performance, based on multiple metrics, confirms that our approach is robust and effective. Future work will focus on integrating more advanced natural language processing techniques, expanding the dataset for greater diversity, and experimenting with ensemble learning strategies to further reduce classification errors. These improvements aim to extend the system's applicability in real-time spam filtering applications.

References

[1] D. Conway and J. M. White, Machine Learning for Email: Spam Filtering and Priority Inbox. Shroff/O'Reilly, 2012.

[2] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2023.

[3] Y. Bouarara, "Machine Learning Techniques in Spam Detection," in Advanced Deep Learning Applications in Big Data Analytics, IGI Global, 2020, pp. 45–67.