Building a Smarter AI-Powered Spam Classifier

Phase-2 Document Submission

Project Overview

Project Name: Building a Smarter AI-Powered Spam Classifier

Project Phase: Phase 2 – Innovation

Phase Overview

In this phase, we'll explore innovative techniques and approaches to building our spam classifier.

One innovative technique we can explore is using pre-trained language models like BERT for feature extraction. These models have demonstrated superior performance in NLP tasks.

Key Objectives

- ❖ **Better Accuracy:** BERT helps classify spam emails more accurately due to its understanding of context
- ❖ **Less Manual Work:** It reduces the need for manual feature engineering, saving time.
- ❖ **Adaptable:** BERT adapts to changing spam tactics.
- ❖ **Interpretable:** Can be combined with techniques for understanding classification decisions.
- ❖ **Scalable:** Works well with large email datasets.
- ❖ **State-of-the-Art:** Aims for top-notch spam classification performance.

Innovation Process

1.Model Selection

BERT as the Feature Extraction Method: BERT, a state-of-the-art pre-trained language model, has shown remarkable success in capturing contextual information from text. Leveraging BERT for feature extraction can significantly enhance our classifier's understanding of email content.

Logistic Regression as a Classifier: Logistic regression is chosen as the classifier to work with BERT embeddings because of its simplicity and efficiency. It complements the complex feature extraction power of BERT.

2.Data Collection and Preparation

Collecting a Diverse Email Dataset: To train a spam classifier effectively, we need a diverse and representative dataset of email texts. This dataset should include various types of spam and legitimate emails to ensure the model's ability to generalize.

Data Preprocessing: Before feeding the data into the model, data preprocessing is essential. This includes tasks like removing HTML tags, extracting relevant features (e.g., sender information, subject lines), and converting text into a suitable format for analysis. Proper label encoding ensures that the data is ready for training.

3. Feature Extraction with BERT

BERT Transformation: BERT transforms email text into high-dimensional embeddings that capture the contextual meaning of words and phrases. These embeddings are rich in information and can significantly improve the model's understanding of email content.

Fine-tuning Consideration: While BERT provides pre-trained embeddings, fine-tuning BERT on the specific task of spam classification may be explored to further optimize its performance for this task.

4.Model Training

Training the Classifier: The training phase involves feeding the model with labeled data, which consists of email texts and their corresponding spam or non-spam labels. The model learns to make predictions based on the extracted features (BERT embeddings) and adjusts its parameters to minimize prediction errors.

Hyperparameter Optimization: Hyperparameter tuning involves experimenting with different settings, such as learning rates or batch sizes, to find the configuration that leads to the best model performance.

5.Evaluation and Validation

Performance Metrics: To measure the effectiveness of the spam classifier, we use standard performance metrics such as accuracy, precision (correctly classified spam emails), recall (spam emails correctly identified), F1-score (a balance of precision and recall), and ROC curves (receiver operating characteristic).

Validation Methods: The model's generalization ability is validated either through cross-validation (dividing the data into multiple subsets for training and testing) or using a hold-out dataset that the model has never seen before.

6.Interpretability

Understanding Model Decisions: Implementing interpretability techniques such as attention mechanisms or feature importance analysis helps us gain insights into why the model makes certain spam/ham classifications. This transparency is essential for trust and debugging.

7. Deployment

Production Deployment: Deploying the trained spam classifier in a production environment, such as an email server or filtering system, ensures that users benefit from its capabilities. It's a critical step in realizing the impact of the project.

Documentation

1.Model Selection

Feature Extraction: Chose BERT as the pre-trained language model for feature extraction.

Classifier Choice: Selected logistic regression as the classifier for working with BERT embeddings.

2.Feature Extraction with BERT

BERT Transformation: Utilized pre-trained BERT to convert email text into dense vector representations (embeddings).

Fine-tuning Consideration: Explored the possibility of fine-tuning BERT for better task-specific performance.

3. Model Training

Classifier Training: Trained the spam classifier using BERT embeddings and labeled data.

Hyperparameter Optimization: Conducted experiments to optimize hyperparameters for improved performance.

4.Evaluation and Validation

Performance Metrics: Evaluated the classifier's performance using standard metrics like accuracy, precision, recall, F1-score, and ROC curves.

Validation Methods: Validated the model against hold-out data or through cross-validation.

5.Deployment

Production Deployment: Deployed the trained spam classifier within an email server or filtering system.

Conclusion

In conclusion, leveraging pre-trained language models such as BERT for feature extraction is a powerful and innovative approach to improve the performance of natural language processing tasks, enabling better understanding and utilization of textual data.