Email Spam Classification Project Documentation

Introduction

This project implements a deep learning model to classify emails as either spam or legitimate (ham). Email spam filtering is an essential application of natural language processing and machine learning that helps protect users from unwanted, potentially harmful messages. The system uses advanced text processing techniques and deep learning architecture to achieve high accuracy in classification.

Project Overview

The goal of this project is to build a robust email spam classifier using natural language processing techniques and deep learning. The key steps in the project include:

- Loading and exploring the dataset
- Preprocessing the text data
- Balancing the dataset
- Building and training a deep learning model with LSTM architecture
- Evaluating the model's performance

Dataset Description

The dataset used for this project contains labeled emails classified as either "spam" or "ham" (legitimate emails). The dataset has the following characteristics:

- Total number of samples: 5,171
- Number of features: 4
- Key features: text content and classification labels
- Class distribution: Imbalanced, with more ham emails than spam emails

Methodology

Data Preprocessing

Several preprocessing steps were applied to the text data to improve model performance:

1. **Text Cleaning**:
    - Removal of the word "Subject" from all emails
    - Removal of punctuation marks
    - Removal of common English stopwords
2. **Dataset Balancing**:
    - Downsampling the majority class (ham) to match the number of spam emails
    - This step ensures the model learns equally from both classes
3. **Text Tokenization**:
    - Converting words to numerical tokens
    - Padding sequences to ensure uniform length

Model Architecture

The model uses a sequential architecture with the following layers:

1. **Embedding Layer**: Converts token IDs to dense vectors of fixed size
    - Input dimension: Size of vocabulary + 1
    - Output dimension: 32
    - Input length: 100 (maximum sequence length)
2. **LSTM Layer**: Processes sequential information
    - Output dimension: 16 units
3. **Dense Layer**: Extracts relevant features
    - 32 units with ReLU activation
4. **Output Layer**: Produces classification result
    - 1 unit with sigmoid activation for binary classification

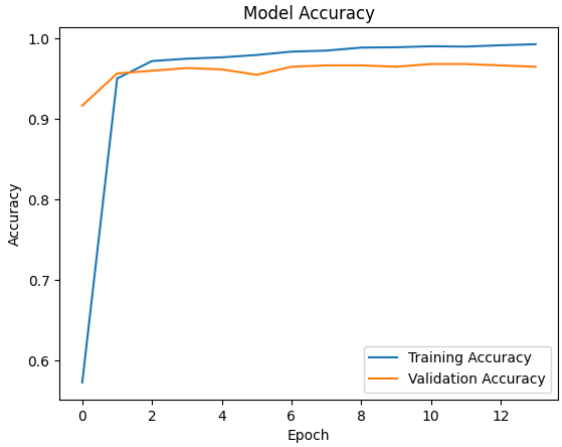
Results and Evaluation

The model achieved excellent performance metrics:

- **Test Loss**: 0.1202
- **Test Accuracy**: 97.00%

These results indicate that the model is highly effective at distinguishing between spam and legitimate emails after training.

The training and validation accuracy curves show how the model's performance improved throughout the training process:



The visualization indicates that the model achieved stable and high accuracy without overfitting to the training data.

Conclusion

This project successfully implemented an email spam classifier using deep learning techniques. The LSTM-based model achieved 97% accuracy on the test set, demonstrating its effectiveness in distinguishing between spam and legitimate emails.

Key achievements:

- Successfully balanced the dataset to improve training
- Applied appropriate text preprocessing techniques
- Implemented an effective deep learning architecture
- Achieved high classification accuracy

Future Improvements

Several potential enhancements could further improve the model's performance:

1. **Feature Engineering**: Incorporate additional features such as sender information, email structure, and metadata.
2. **Model Architecture**: Experiment with bidirectional LSTMs, attention mechanisms, or transformer-based models.
3. **Hyperparameter Tuning**: Use techniques like grid search or Bayesian optimization to find optimal hyperparameters.
4. **Data Augmentation**: Generate synthetic examples to increase the diversity of the training data.
5. **Transfer Learning**: Leverage pre-trained language models like BERT or GPT for improved text representation.
6. **Real-world Deployment**: Implement the model as part of an email filtering system with continuous learning capabilities.