Fake Email Classifier Model Documentation

Yusup Orazov

February 28, 2024

Abstract

This document provides detailed documentation of the fake email classifier model implementation. The model is built using PyTorch and uses a LSTM network architecture for text classification. Key aspects covered include dataset preparation, model definition, training loop, evaluation methodology, and results.

1 Introduction

The spread of fake and phishing emails has become a major issue in recent times. This project aims to build an automated classifier that can effectively distinguish between fake/phishing emails and legitimate emails. The model is implemented in PyTorch using a LSTM neural network, and is evaluated using k-fold cross validation on a manually prepared dataset.

2 Dataset

The dataset consists of 150 fake/phishing email texts collected manually. The texts span a diverse range of topics that aim to appear as legitimate notifications and requests to trick the user. Each text is labeled as 0 to indicate fake email. The dataset is split into train/validation by a 90/10 ratio for model development and evaluation.

3 Preprocessing

The email texts are tokenized using BERT tokenizer from HuggingFace Transformers library. This provides standardized word token embeddings as input to the LSTM model. The tokenized texts are padded to equal lengths for creating mini-batches.

4 Model Architecture

The neural network model implementing the classifier consists of the following components:

- **Input Layer:** Tokenized email text embeddings of length L
- **LSTM Layer:** Single LSTM layer with 80 hidden units
- **Fully-Connected Layer:** Linear layer with 2 output units and softmax activation for binary fake/real classification
- **Loss function:** Cross-entropy loss
- **Optimizer:** Adam

5 Training Methodology

The model is trained for 20 epochs with early stopping using k-fold cross validation with $k = 3$. The hyperparameters used are:

- Batch size = 32
- Learning rate = 0.00024
- Training set shuffle = True

The average validation accuracy and loss across folds are computed to evaluate model performance.

6 Results

The model achieves an average validation accuracy of 95.82

7 Conclusion

The fake email classifier model provides excellent accuracy in identifying fraudulent emails. Next steps could involve testing it against more sophisticated phishing emails and deploying the trained model in applications for spam/threat detection. The model can be further improved by using larger datasets, hyperparameter tuning, and recent advancements in NLP architectures.