Spam Email Detection

Hassan Said Mohamed : Word 2Vec

Supervisor: Dr. Wesam Ahmed

Group Number: 6

Submission date: May 12th, 2025

1. Introduction

This project implements a spam message classification system using both classical Machine Learning models and a deep learning LSTM model. The dataset used is the popular "spam.csv" dataset.

2. Data Loading and Preprocessing

The dataset is loaded using pandas.

Only relevant columns 'v1' (label) and 'v2' (text) are used.

Labels are converted to binary (ham=0, spam=1).

Text is preprocessed by converting to lowercase.

3. Classical Machine Learning Models

TF-IDF vectorization is applied to convert text data to numeric features.

The dataset is split into training and test sets (80/20 split) with stratification.

Three ML models are trained:

Logistic Regression

Random Forest

Support Vector Machine (SVM)

Models are evaluated using accuracy, precision, recall, and F1-score.

Confusion matrices are generated for each model

4. Word2Vec Embeddings and LSTM Model

Tokenization is performed on the text data.

A Word2Vec model is trained on the tokenized text.

Sequences are padded to a fixed length.

An embedding matrix is created from the Word2Vec embeddings.

A Bidirectional LSTM model is defined with embedding, LSTM, dropout, and dense layers.

The model is compiled with binary crossentropy loss and Adam optimizer.

Early stopping is used to prevent overfitting.

The model is trained on the training sequences with class weights to balance classes.

Training and validation accuracy and loss are monitored.

5. Results

Classical ML models achieved high accuracy with Random Forest and SVM outperforming Logistic Regression.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Recall | Prescision | F1-measure |
| Logistic regression | 99.88% | 97.76% | **0.98(Spam), 0.95(Ham)** | **0.99(Spam), 0.89(Ham)** | |  |  | | --- | --- | | **0.99 (Spam), 0.92 (Ham)** |  | |
| (SVM): | 99.86% | 95.43% | **0.99(Spam),**  **0.72(Ham)** | **0.96(Spam)**  **,**  **0.92(Ham)** | **0.97**  **(Spam), 0.81 (Ham)** |
| Random Forest | 96.27% | 97.31% | **1.0**  **(Spam),**  **0.82**  **(Ham)** | **0.97(Spam), 0.98 (Ham)** | **0.98**  **(Spam), 0.89 (Ham)** |
| LSTM | 100.00% | 98.77% | **1.0**  **(Spam),**  **0.93 (Ham)** | **0.99**  **(Spam),**  **0.97 (Ham)** | **0.99**  **(Spam), 0.95 (Ham)** |

The LSTM model showed strong performance with validation accuracy exceeding 98%.

Confusion matrices illustrate the model predictions versus true labels for both ML and LSTM models.

6. Conclusion

This project demonstrates the effectiveness of combining classical machine learning with deep learning techniques for spam detection. Word embeddings and LSTM provide a powerful method to capture sequential dependencies in text, improving classification accuracy.

7. Future Work

Experiment with pre-trained embeddings like GloVe or FastText.

Tune hyperparameters of LSTM and ML models.

Add more preprocessing steps like stopword removal and stemming.

Explore transformer models like BERT for further improvements.

References