Spam Email Detection

Mahmoud helal amen : TF-IDF feature extraction

Supervisor: Dr. Wesam Ahmed

Group Number: 6

Submission date: May 12th, 2025

Chapter 1

Introduction

This project implements a spam detection system using various machine learning models and deep learning models. The dataset is processed through multiple steps, including text preprocessing, transformation, and classification using different algorithms such as Logistic Regression, Naïve Bayes, Random Forest, and LSTM.

Chapter 2

Code Implementation

Preprocessing:

The preprocessing steps

1. **Loading Dataset:**

The data is read from a CSV file and unnecessary columns are dropped and used columns are renemed to message and category instead of v1 and v2.

**2. Cleaning Data:**

- Removing missing values and duplicates.

- Converting text to lowercase.

- Removing special characters, numbers, and punctuation.

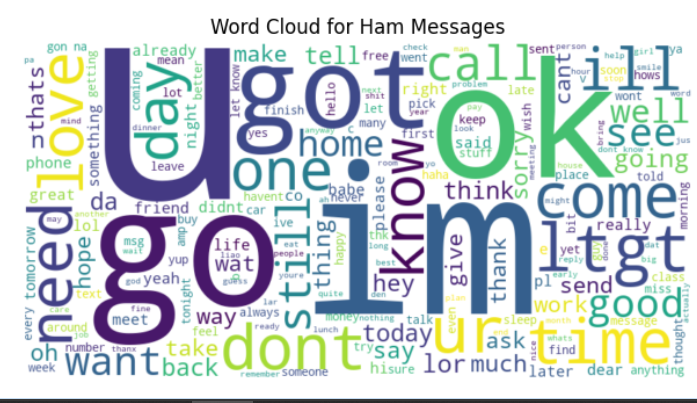
- Removing stop words.

- Tokenization.

- Lemmatization.

- Mapping categories (spam: 1, ham: 0)

Visualization:





**Text Transformation:**

We used TF IDF to transform the text into numirical value based on importance.

**Models:**

**Machine learning models:**

**1-logistic regression**

**2-Naïve Bayes**

**3-Random forest**

**Deep learning models:**

**1-LSTM**

**TF IDF with machine and deep learning models:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training Accuracy | Testing Accuracy | Accuracy | Recall | Prescision | F1-measure | AUC Value |
| Logistic regression | 96.08% | 94.29% | 94.29% | |  |  | | --- | --- | | **0.64 (Spam), 0.99 (Ham)** |  | | |  |  | | --- | --- | | **0.93 (Spam), 0.94 (Ham)** |  | | |  |  | | --- | --- | | **0.76 (Spam), 0.97 (Ham)** |  | | 0.98 |
| Naive Bayes | 96.86% | 96.32% | 96.32% | **0.74 (Spam), 1.00 (Ham)** | **1.00**  **(Spam), 0.96 (Ham)** | **0.85**  **(Spam), 0.98 (Ham)** | 0.98 |
| Random Forest | 100% | 96.71% | 96.71% | **0.77**  **(Spam), 1.00 (Ham)** | **1.00**  **(Spam), 0.96 (Ham)** | **0.87**  **(Spam), 0.98 (Ham)** | 0.99 |
| LSTM | 87.71% | 85.98% | 85.98% | **0**  **(Spam), 1.00 (Ham)** | **0**  **(Spam), 0.86 (Ham)** | **0**  **(Spam), 0.92 (Ham)** | 0.99 |

Figure 2.2: Logistic regression with TF IDF confusion matrix and Roc Curve

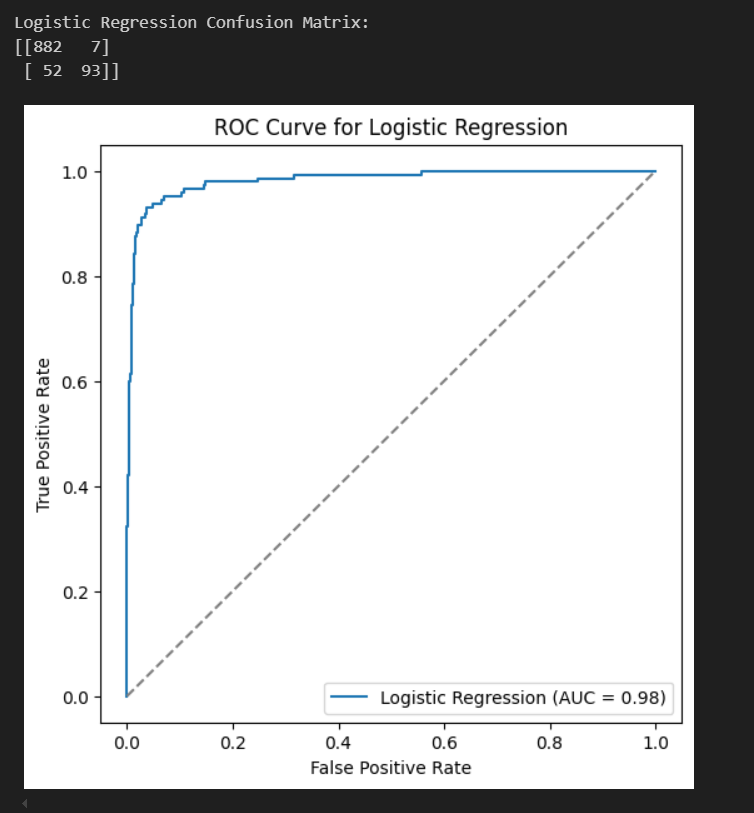


Figure 2.3: Naïve Bayes with TF IDF confusion matrix and Roc Curve

A screen shot of a graph

AI-generated content may be incorrect.

Figure 2.4: Random Forest with TF IDF confusion matrix and Roc Curve

A screen shot of a graph

AI-generated content may be incorrect.

Figure 2.5: LSTM with TF IDF confusion matrix and Roc Curve

A screen shot of a graph

AI-generated content may be incorrect.