Email Spam Detection

Ahmed Saad Hamed : GloVe Embedding 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 and deep learning models. The dataset is processed through comprehensive text preprocessing steps, followed by word embedding using GloVe. Classification is then performed using Logistic Regression, Random Forest, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) models.

Chapter 2

Code Implementation

Preprocessing:

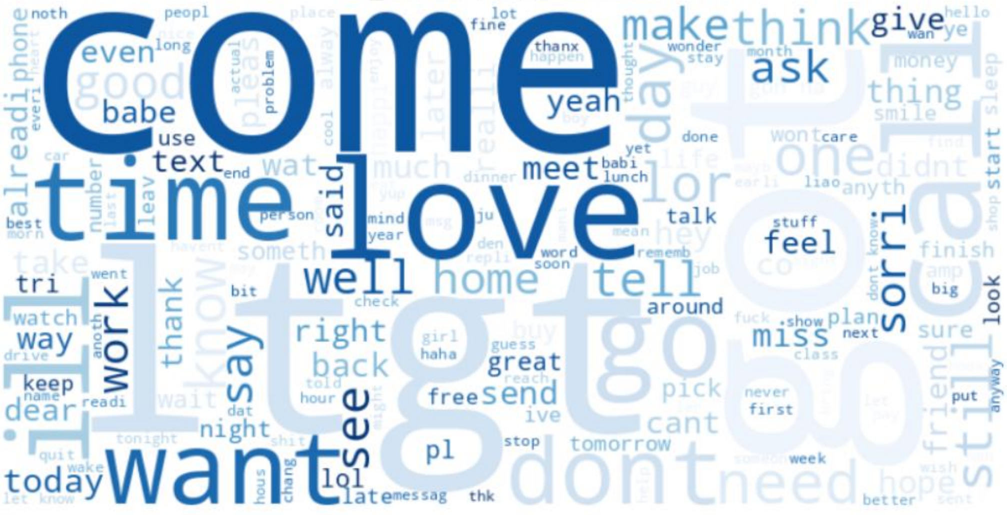
The preprocessing steps:

- Loading Dataset:
  - Cleaning Data:
  - Removing missing values and duplicates.
  - Converting all text to lowercase.
  - Removing special characters, numbers, and punctuation.

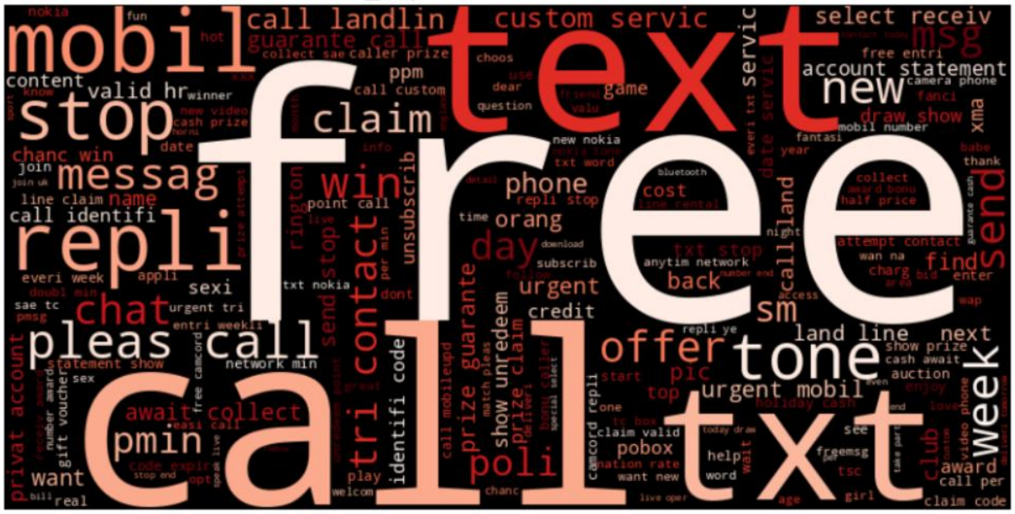
- Removing stop words.
- Tokenization using NLTK.
- Lemmatization using WordNetLemmatizer.
- Mapping categories (

Visualization:

□ Ham Word Cloud



□ Spam Word Cloud



Text Transformation:

We used GloVe (Global Vectors for Word Representation) to generate semantic vector embeddings for each message based on pretrained 100-dimensional embeddings.

Models:

Machine Learning Models:

1. Logistic Regression
2. Random Forest
3. Support Vector Machine (SVM)

Deep Learning Model:

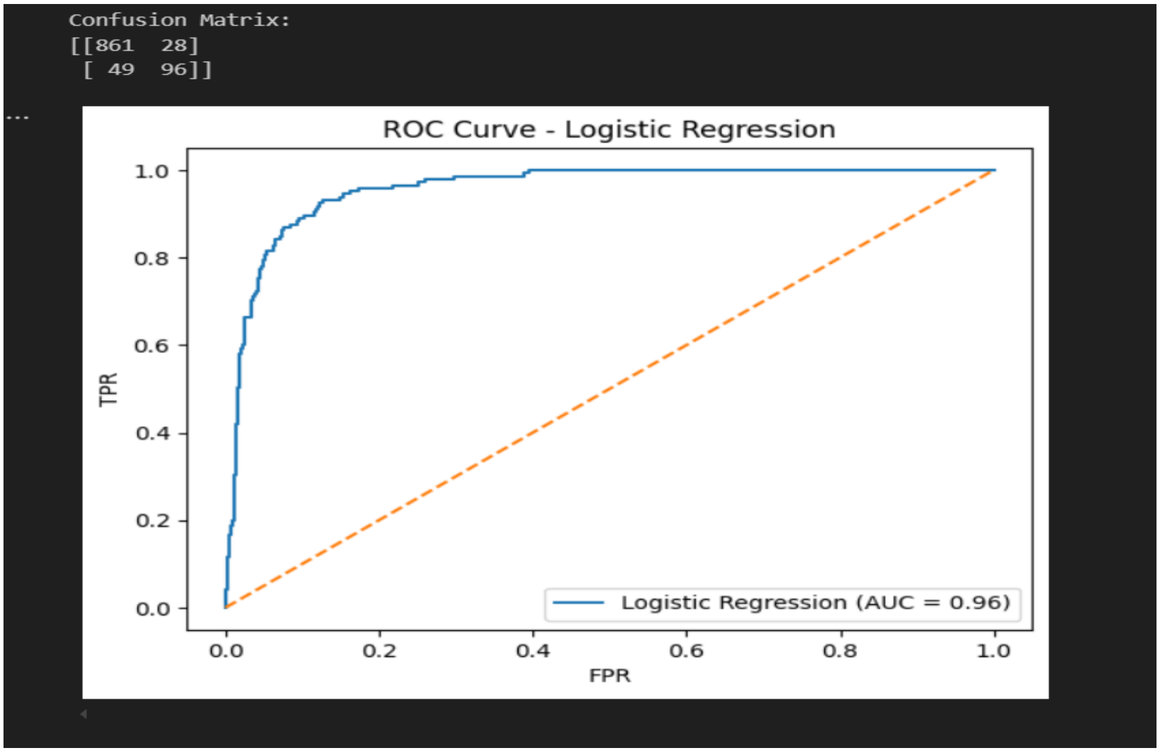
1. LSTM (Long Short-Term Memory)

GloVe Embedding with Machine and Deep Learning Models:

Model	Training Accuracy	Testing Accuracy	Accuracy	Recall	Precision	F1-measure	AUC Value
Logistic regression	93.37%	92.55%	92.55%	**0.66(Spam), 0.97 (Ham)**	**0.77(Spam), 0.95 (Ham)**	**0.71(Spam), 0.96 (Ham)**	0.96

Random Forest	99.93%	94.39%	94.39%	0.61 (Spam), 1.00 (Ham)	0.98(Spam), 0.94 (Ham)	0.75(Spam), 0.95 (Ham)	0.97
SVM	97%	95.94%	94.39%	0.79(Spam), 0.99 (Ham)	0.91(Spam), 0.97 (Ham)	0.85(Spam), 0.95 (Ham)	0.98
LSTM	87.71%	85.98%	85.98%	0(Spam), 1.00 (Ham)	0(Spam), 0.86 (Ham)	0(Spam), 0.92 (Ham)	0.68

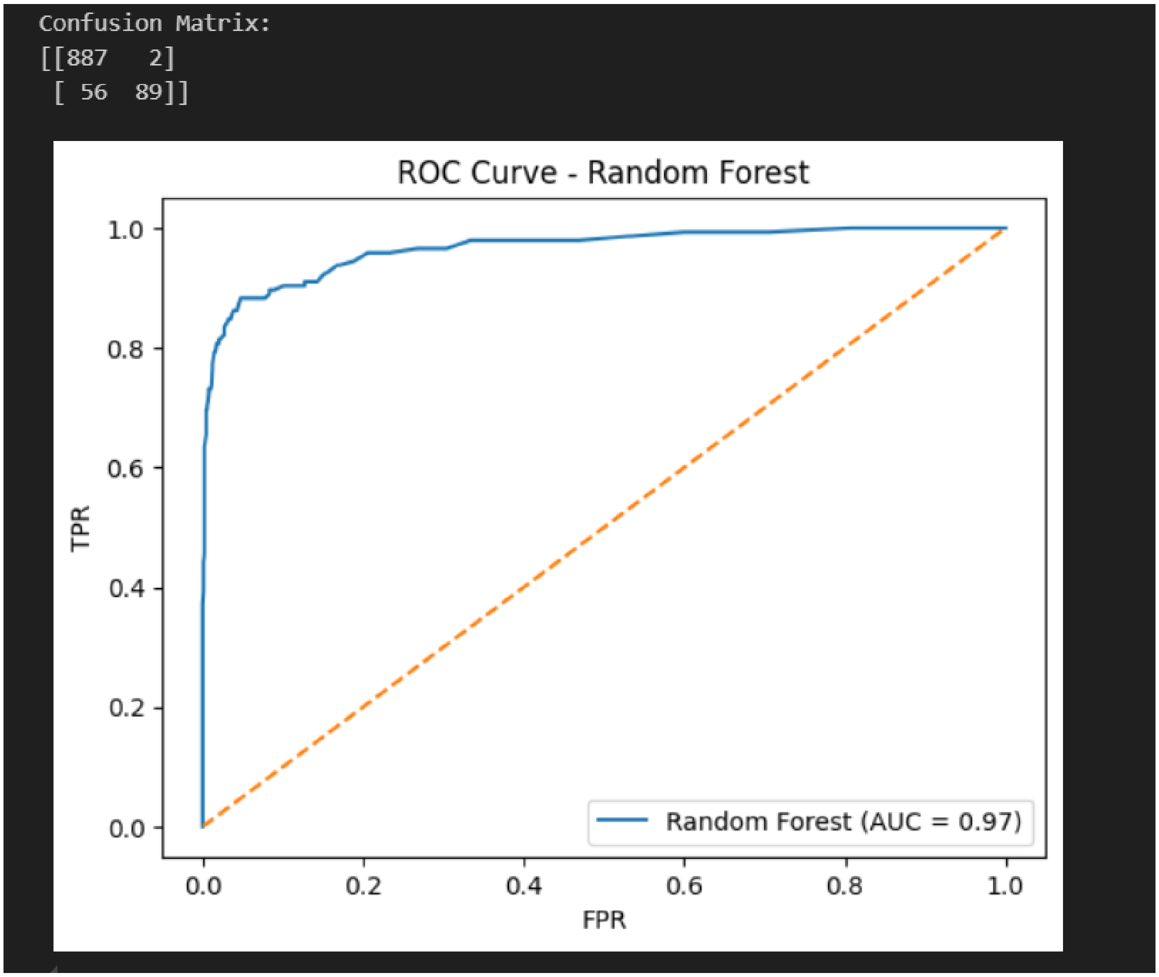
Figure 2.2: Logistic Regression with GloVe - Confusion Matrix and ROC Curve



```
Confusion Matrix:
[[861  28]
 [ 49  96]]
```

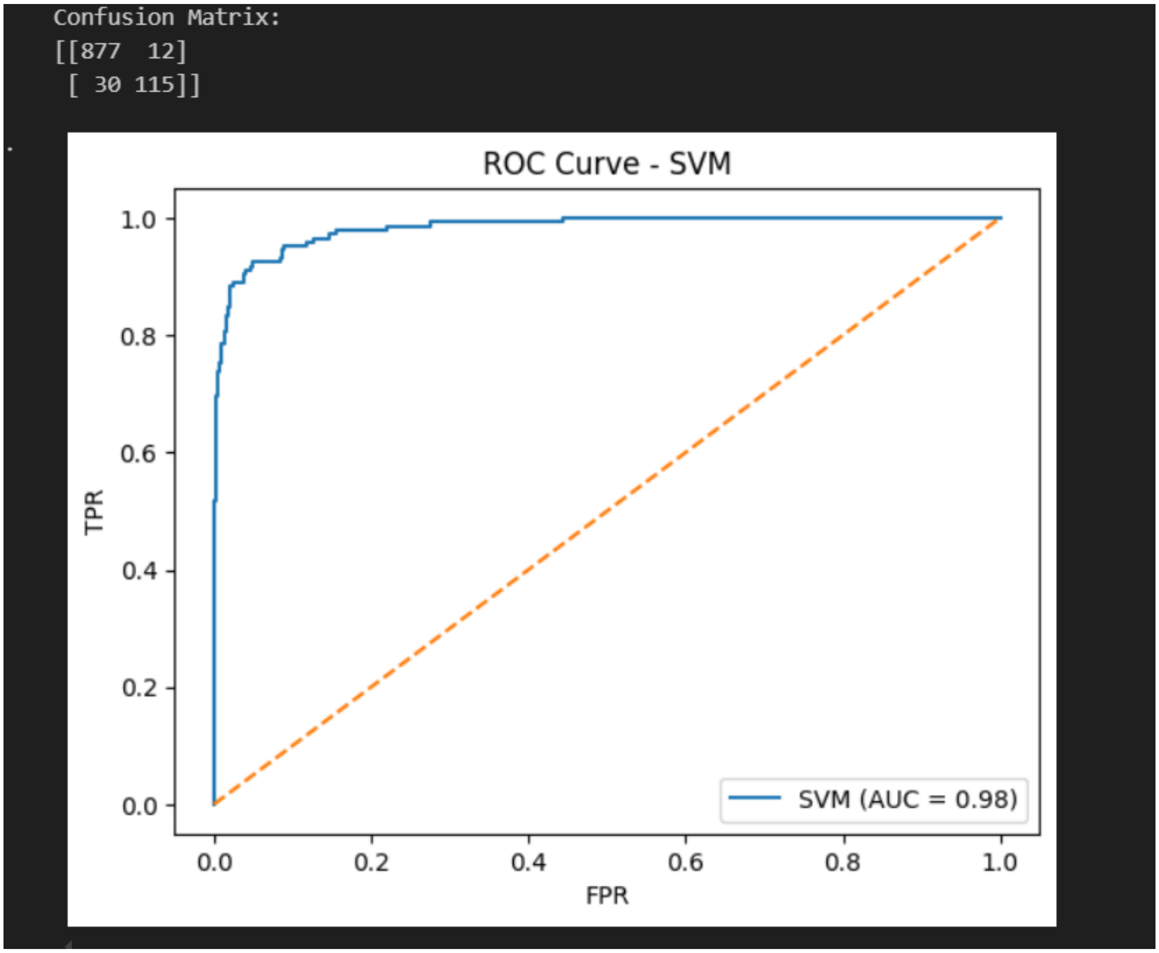
...
ROC Curve - Logistic Regression
1.0
0.8
0.6
0.4
0.2
0.0
TPR
0.0
0.2
0.4
0.6
0.8
1.0
FPR
Logistic Regression (AUC = 0.96)

Figure 2.3: Random Forest with GloVe - Confusion Matrix and ROC Curve



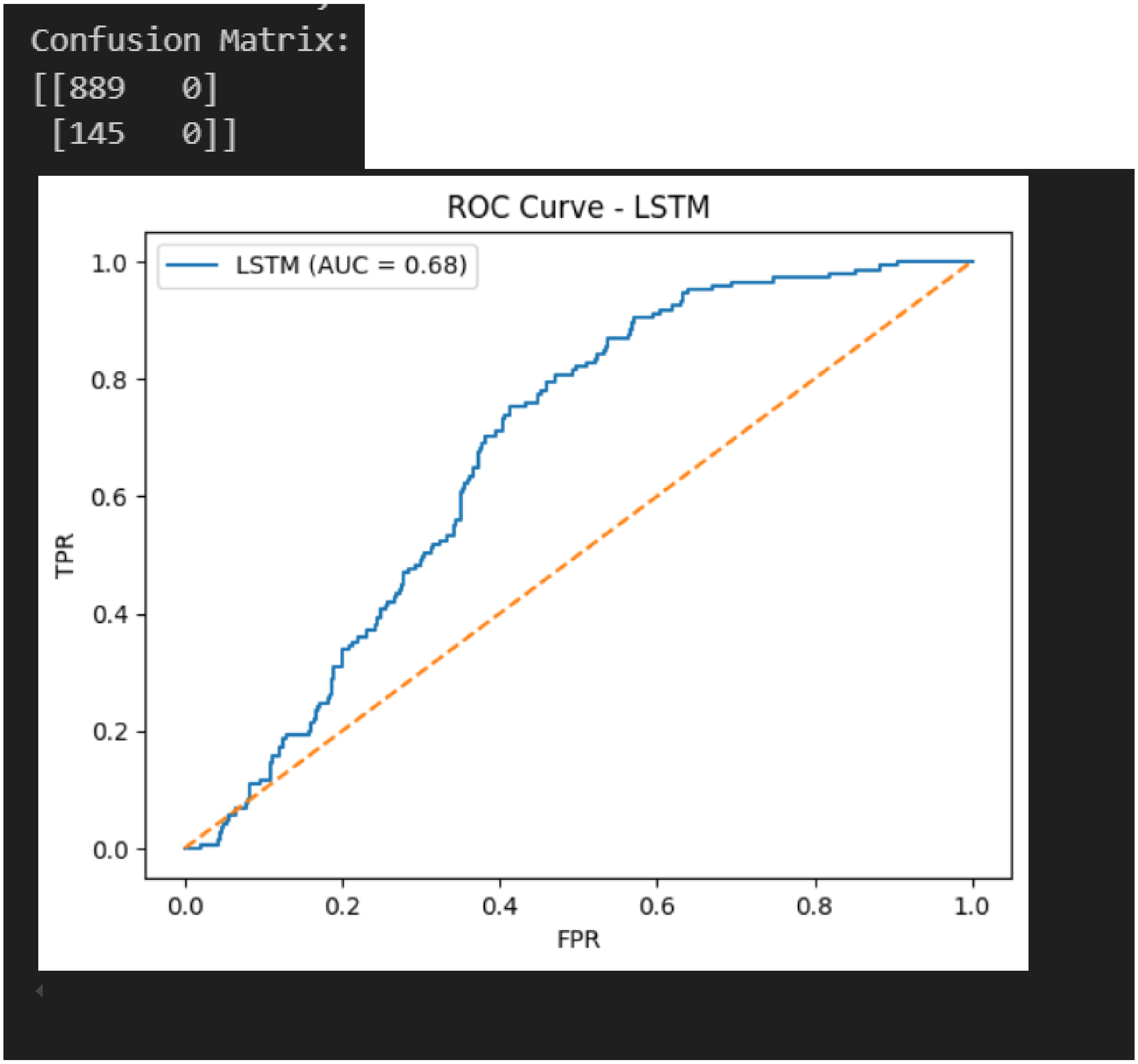
```
Confusion Matrix:
[[887   2]
 [ 56  89]]
```

Figure 2.4: SVM with GloVe - Confusion Matrix and ROC Curve



```
Confusion Matrix:
[[877  12]
 [ 30 115]]
```

Figure 2.5: LSTM with GloVe - Confusion Matrix and ROC Curve



```
Confusion Matrix:
[[889   0]
 [145   0]]
```