Spam\_Email Detection

Ahmed Saad Hamed : FastText Embedding feature extraction

Supervisor: Dr. Wesam Ahmed

Group Number: 6

Submission date: May 12th, 2025

Introduction

Spam emails are unsolicited messages that flood user inboxes and often carry malicious content or advertisements. Filtering such messages is critical for cybersecurity and user experience. This project presents a spam email classification system using both traditional machine learning models and a deep learning approach, leveraging FastText word embeddings for rich semantic representation.

Chapter 2

Preprocessing

Text preprocessing is crucial for converting raw text into a usable form for models. The steps followed include:

* - Contraction expansion using the contractions library
* - Lowercasing and punctuation removal
* - Digit removal and stopword filtering
* - Lemmatization using WordNetLemmatizer

Embedding Generation

A FastText embedding model was trained using the preprocessed tokenized corpus. Sentence-level vectors were generated by averaging the embeddings of each token.

Visualization:

Figure 2.1: Word cloud

Text Transformation:

We Used FastText Embeddings technique

it was applied:

* Each text message was tokenized and preprocessed.
* FastText was trained on these tokens to build word embeddings.
* For each message, the embedding vectors of words were averaged to produce a fixed-size sentence vector (100 dimensions).

Machine Learning Models

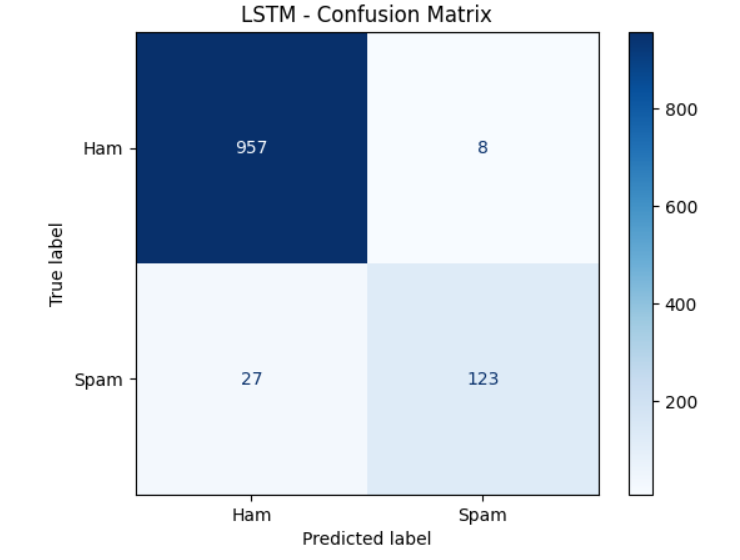
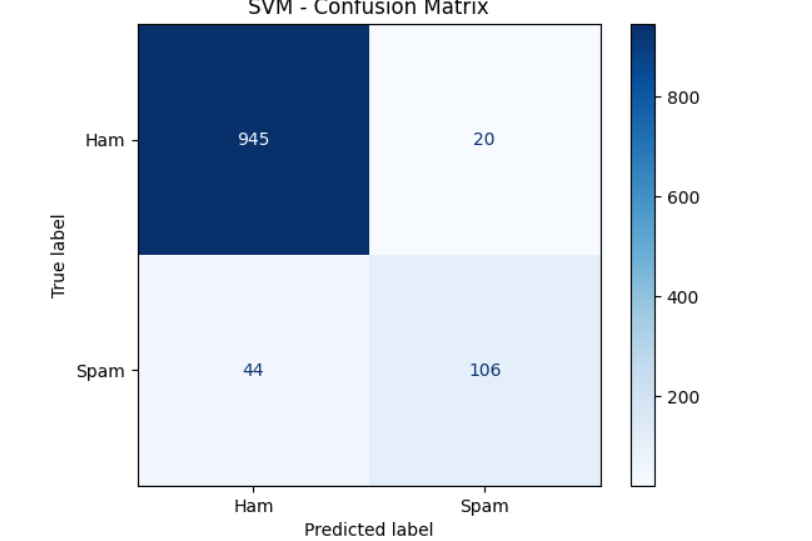
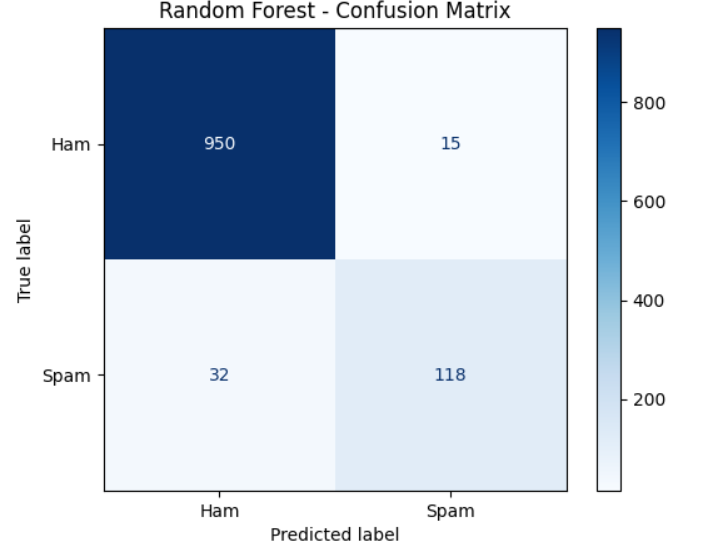
-We trained and tuned multiple ML models using FastText embeddings as features:  
- Logistic Regression  
- Support Vector Machine (SVM)  
- Random Forest  
Each model was evaluated using Accuracy, Precision, Recall, F1-Score, and AUC Score. GridSearchCV was used for hyperparameter optimization with 5-fold cross-validation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Models | Training Accuracy | Testing Accuracy | Accuracy | Recall | Precision | F1-measure | AUC Value |
| Logistic Regression (FastText) | 0.97 | 0.96 | |  | | --- | |  |  |  | | --- | |  |   0.96 | 0.95 | 0.97 | 0.96 | |  | | --- | |  |  |  | | --- | | 0.98 | |
| Random Forest (FastText) | 0.99 | 0.97 | 0.97 | 0.95 | 0.94 | 9.94 | 0.96 |
| SVM | |  | | --- | |  |  |  | | --- | | 0.98 | | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.97 |
| LSTM | |  | | --- | |  |  |  | | --- | | 0.97 | | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.97 |

**Visualization**

1. **Confusing MatrixesA diagram of a logistic regression

   AI-generated content may be incorrect.**

****

**Deep Learning**

An LSTM-based model was implemented using a FastText-initialized embedding layer.  
Architecture:  
- Embedding Layer (FastText, frozen)  
- LSTM (64 units)  
- Dropout (0.5)  
- Dense (1 neuron, sigmoid)

Figure 5.1: LSTM ROC Curve