Phishing Detection

Using NLP

Report

November 7, 2024

1 Introduction

Project Overview

- **Objective**: The goal of this project is to classify phishing emails using machine learning algorithms and compare the performance of **Logistic Regression** and **Random Forest Classifier**.
- **Dataset**: A dataset containing labeled emails (phishing or non-phishing) with various features was used for training and testing.

2 Data Preprocessing

Feature Engineering

- **Feature Extraction**: Text features such as subject and body content were transformed into numerical data using TF-IDF vectorization.
- **Additional Features**: Features like the number of URLs, sentiment scores, email domain types, etc., were also included.

3 Machine Learning Models

Model Architecture

- **Logistic Regression**: A linear model that estimates the probability that a given input belongs to a certain class (phishing or non-phishing). It assumes a linear relationship between features.
- **Random Forest Classifier**: An ensemble learning method that constructs multiple decision trees and aggregates their results to improve accuracy and reduce overfitting.

4 Model Training and Evaluation

Evaluation Framework

- **Training**: Both models were trained on the preprocessed data.
- **Evaluation Metrics**:
  - **Accuracy**: The percentage of correctly classified emails.
  - **Precision**: The proportion of phishing emails correctly identified among all predicted phishing emails.
  - **Recall**: The proportion of actual phishing emails that were correctly classified.
  - **F1-Score**: The harmonic mean of precision and recall.

5 Results Comparison

Metric	Logistic Regression	Random Forest Classifier
Accuracy	0.9938	0.9942
Precision	0.99	1.00
Recall	1.00	0.99
F1-Score	0.99	0.99

Table 1: Performance Comparison of Models

5.1 Logistic Regression Results

Confusion Matrix

```
[[5157   41]
 [  32 6517]]
```

Classification Report

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5198
           1       0.99      1.00      0.99      6549

    accuracy                           0.99     11747
   macro avg       0.99      0.99      0.99     11747
weighted avg       0.99      0.99      0.99     11747
```

5.2 Random Forest Classifier Results

Confusion Matrix

```
[[5167   31]
 [  37 6512]]
```

Classification Report

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5198
           1       1.00      0.99      0.99      6549

    accuracy                           0.99     11747
   macro avg       0.99      0.99      0.99     11747
weighted avg       0.99      0.99      0.99     11747
```

Analysis

Both models performed comparably in terms of accuracy, with Random Forest showing a slight edge due to its ability to model non-linear relationships. Random Forest generally performs better in precision, indicating a lower false positive rate, while both models achieved similar F1-scores, effectively balancing precision and recall.

6 Conclusion

Key Findings

- **Logistic Regression**: The model is simple, interpretable, and performs well on linearly separable data but struggles with complex feature interactions.
- **Random Forest**: It outperforms Logistic Regression in this context by handling non-linear relationships and reducing overfitting through ensemble learning. Thus, it is a better choice for phishing detection in this project.

7 Future Work

Next Steps

- Tuning hyperparameters (e.g., regularization for Logistic Regression, tree depth for Random Forest).
- Experimenting with more advanced models like **XGBoost** or **Neural Networks** for further performance improvements.
- Incorporating real-time phishing detection mechanisms.