Performance Analysis of Naive Bayes Classifiers for Phishing Email Detection

Author: Rayile Adam

Course: ECE 592A - Career Development

Date: August 1, 2025

Abstract

This report details the implementation and evaluation of a Multinomial Naive Bayes (MNB) classifier for the task of phishing email detection. The model was tested with two distinct text feature extraction methods: Bag-of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF), using both a limited and a full feature set. The results demonstrate that while both approaches yield high accuracy, the TF-IDF model with a full feature vocabulary provides a superior balance of precision and recall, significantly reducing the number of missed threats. The findings confirm that Naive Bayes is a highly effective and computationally efficient model for real-time email security applications.

1. Introduction

Phishing has become a critical threat in cybersecurity, leveraging deceptive emails to compromise individual and organizational security. The development of effective machine learning (ML) models is essential to mitigate these risks. This report focuses on the application of the Naive Bayes algorithm, a probabilistic classifier well-suited for text categorization, to the problem of phishing detection. The primary challenge lies in handling the imbalanced nature of email data, where a model must maximize threat detection (high recall) while minimizing the misclassification of legitimate emails (high precision).

2. Model and Feature Selection

2.1. Model Selection: Multinomial Naive Bayes

The Multinomial Naive Bayes (MNB) classifier was selected for this task. MNB is a probabilistic model that calculates the probability of an email belonging to a class (phishing or legitimate) based on the frequency of words within its content. Its primary strengths are its computational efficiency and strong performance on high-dimensional and sparse text data, such as that generated by BOW and TF-IDF vectorization.

2.2. Feature Selection

Two feature extraction techniques were evaluated:

1. **Bag-of-Words (BOW):** This technique represents text by counting the frequency of words, converting textual information into numerical features.

2. **TF-IDF (Term Frequency-Inverse Document Frequency):** This method weights words by their importance, considering their frequency in an email and their rarity across the entire dataset, which helps highlight distinctive terms.

3. Model Hyper-parameters and Optimization

3.1. Data Splitting

The dataset was divided into an 80% training set and a 20% test set using stratified sampling to preserve the class distribution. The test set remained untouched during training and was used only for the final, unbiased evaluation.

3.2. Model Hyper-parameters

For the Naive Bayes model, the following configurations were evaluated:

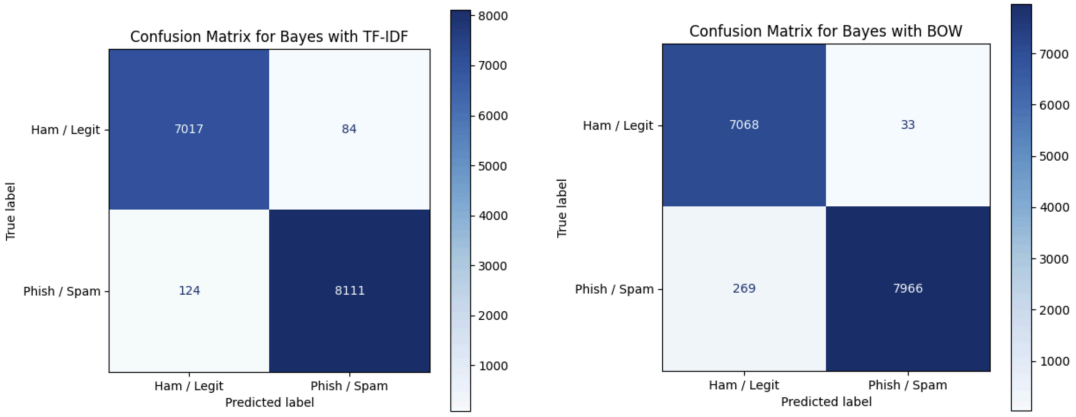
- **Feature Extraction:** BOW and TF-IDF.
- **Vocabulary Size (Maximum Features):** Experiments were run with a limited set of 5,000 features and with the full vocabulary of 1,017,504 features.
- **Smoothing Parameter (alpha):** The model used Laplace smoothing with alpha=1.0 to handle words that may not appear in the training data.

4. Results

The MNB model was evaluated using accuracy, precision, recall, F1-score, and AUC. The comprehensive results are presented in Table I.

Feature	Params	Accuracy	Precision	Recall	F1	AUC	Train & Predict Time
BOW	5000 features Laplace smoothing alpha = 1	0.9538	0.9756	0.9346	0.9547	0.9885	sub sec
TF-IDF	5000 features alpha = 1	0.9598	0.9789	0.9431	0.9607	0.9945	sub sec
BOW	1,017,504 features Laplace smoothing alpha = 1	0.9803	0.9959	0.9673	0.9814	0.9977	32.4s
TF-IDF	1,017,504 features Laplace smoothing alpha = 1	0.9864	0.9897	0.9849	0.9873	0.9991	33.3s

The confusion matrices for the two best-performing models (using the full feature set) provide a detailed breakdown of their predictions on the 15,336 test emails.



5. Discussion

The results from the Naive Bayes experiments reveal a classic trade-off between precision and recall, which is central to any security-focused classification problem.

- The **BOW model** achieved a near-perfect precision of 0.9959, meaning it was extremely unlikely to misclassify a legitimate email as phishing (only 33 false positives). However, this came at the cost of a lower recall, resulting in 269 false negatives—phishing emails that were missed.
- The **TF-IDF model** provided a much better balance for a security application. While its precision was slightly lower, it achieved a significantly higher recall of 0.9849. This reduced the number of false negatives to just 124, representing a **54% decrease in missed threats**.

In phishing detection, a false negative (a missed threat) is far more critical than a false positive (an inconvenience). Therefore, the model that maximizes recall is functionally superior. The TF-IDF model's ability to weigh important words allowed it to better identify threats, making it the more effective and secure choice.

6. Conclusion

The Naive Bayes classifier is a highly effective and efficient model for phishing email detection. The experiments demonstrate that the choice of feature extraction is critical, with the **TF-IDF method providing superior performance** by significantly improving the model's ability to recall and neutralize threats. For a real-world deployment, the Naive Bayes model with TF-IDF vectorization is the recommended approach, as it offers the best combination of accuracy, speed, and, most importantly, security.