# Machine Learning & Data Science II

**Spam Detection using Naive Bayes**


Bachelor's degree course - Mechatronics Design & Innovation

5. Semester

Lecturer: D. T. McGuiness, PhD

Author: Noel Hack


January 2, 2025

# Contents

# 1 Introduction

This project addresses spam detection through a Naive Bayes classifier, focusing on preprocessing, feature extraction with n-grams, and a custom probability threshold to enhance detection accuracy. The goal is to classify emails into ham or spam with high accuracy, achieving an optimal balance between identifying spam and avoiding misclassifications of ham.

# 2 Methodology

## 2.1 DATA LOADING

The dataset comprises three categories of emails:

- `easy_ham`: Non-spam emails with typical structures.
- `hard_ham`: Non-spam emails that may resemble spam.
- `spam`: Spam emails with typical spam characteristics.

Each email is labeled as 0 for ham and 1 for spam, and all emails are combined into a single dataset for training and testing.

## 2.2 DATA PROCESSING

To reduce complexity for the classifier, the following steps are applied:

- Convert text to lowercase to ensure case-insensitive processing.
- Replace numerical values with the token `NUMBER`.
- Replace URLs with the token `URL`.
- Remove non-word characters to clean extraneous symbols and punctuation.

These steps reduce variability in email text representation, aiding the classifier in generalizing across diverse formats.

## 2.3 VECTORIZATION WITH N-GRAMS

In the next step `CountVectorizer` for feature extraction with the following configurations is applied:

- **N-gram Range**: Unigrams and bigrams are used to capture individual words and pairs of words, providing additional context for distinction.
- **Stop Words Removal**: Common English stop words are excluded to retain only informative terms.
- **Max/Min Document Frequency**: Terms appearing in over 50% of emails are discarded to avoid frequent, non-discriminative words.

## 2.4 CLASSIFICATION MODEL: NAIVE BAYES

The classifier used is a Multinomial Naive Bayes model with a smoothing parameter, `alpha`, set to 0.001. This small `alpha` value enables the model to focus on low-frequency terms that may indicate spam, improving sensitivity to rare but meaningful patterns.

# 3 Results and Evaluation

## 3.1 ACCURACY AND CLASSIFICATION REPORT

The classifier achieves an accuracy of 99.6%, which demonstrates its strong performance in spam detection. Table 3.1 presents precision, recall, and F1-score, offering a detailed view of model performance across both spam and ham classes.

Table 3.1: Classification Report

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| Ham | 0.999 | 0.995 | 0.997 |
| Spam | 0.989 | 0.997 | 0.993 |

## 3.2 PROBABILITY DISTRIBUTION OF SPAM PREDICTION

The histogram in Figure 3.1 illustrates the distribution of predicted spam probabilities, showing a pronounced concentration of values near 0 and 1. This clustering reflects the classifier's high confidence, with most emails being classified with near certainty as either ham or spam. The red line denotes the spam classification threshold set at 0.9, intended to help refine the balance between false positives and false negatives. However, due to the strong polarization of probabilities around 0 and 1, this threshold adjustment has minimal impact on classification outcomes.

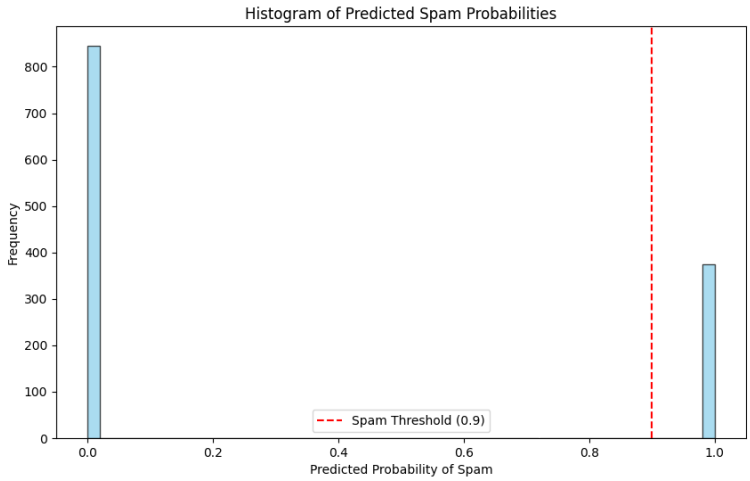


Figure 3.1: Histogram of Predicted Spam Probabilities

### 3.3 CONFUSION MATRIX

Figure 3.2 presents the confusion matrix, displaying correct and incorrect classifications for ham and spam emails. The matrix highlights the model's capability to accurately distinguish between spam and ham emails.

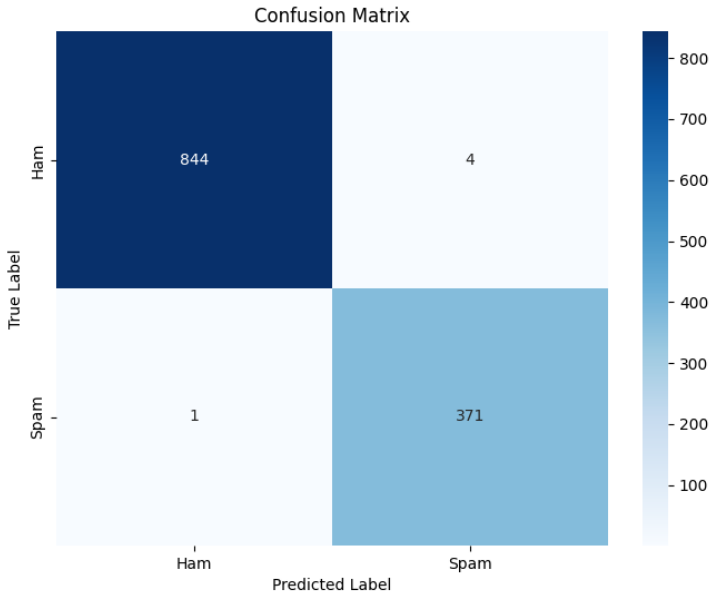


Figure 3.2: Confusion Matrix for Spam Detection

# 4 Conclusion

The Naive Bayes classifier, augmented with n-gram vectorization and threshold adjustments, achieves robust spam detection with an accuracy of 99.6%. N-grams capture contextual relationships in text, improving classification of complex spam.

# List of Figures

# List of Tables