# SMS Spam Detection Project

## 1. Introduction

This project focuses on building a spam classifier using Natural Language Processing (NLP) and machine learning techniques. It uses the "SMSSpamCollection" dataset, which contains labeled SMS messages categorized as either "ham" (not spam) or "spam".

## 2. Dataset

- **Structure**: The dataset has two columns:
    - label: Message type (ham or spam)
    - message: The actual text message

## 3. Exploratory Data Analysis

- Dataset Size: (5572, 2)
- Class Distribution:
    - Ham: 4825
    - Spam: 747

```
Dataset Size: (5572, 2)

Class Distribution:
 ham     4825
spam     747
Name: label, dtype: int64
```

- A length column was added to study message lengths.
- Visualizations:
    - Class distribution bar plot
    - Histogram of message lengths

## 4. Text Preprocessing

Each message undergoes the following steps:

- Lowercasing

- Removal of punctuation and numbers
- Tokenization
- Stopwords removal using NLTK
- Stemming with PorterStemmer

This is encapsulated in the preprocess_text() function.

**5. Feature Engineering**

- CountVectorizer is used to convert cleaned text into a bag-of-words (BoW) model.
- Target labels are encoded (ham = 0, spam = 1)

**6. Model Training**

- Algorithm: **Multinomial Naive Bayes (MNB)**
- Data split: 80% training, 20% testing
- The model is trained on vectorized messages.

**7. Evaluation**

- **Accuracy**: Around 98% on the test set.
- **Confusion Matrix**:
    - High true positive and true negative rates
    - Low false positives and false negatives

The confusion matrix plot gives a visual overview of prediction performance.

**8. Inference Function**

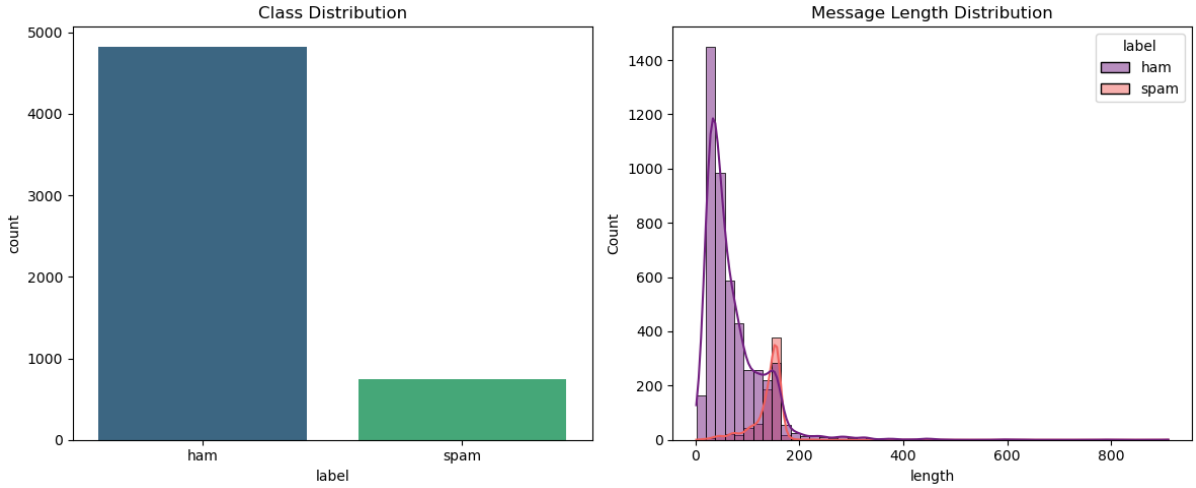
A custom function predict_spam() takes a raw message, processes it, and returns a prediction (Spam or Ham).

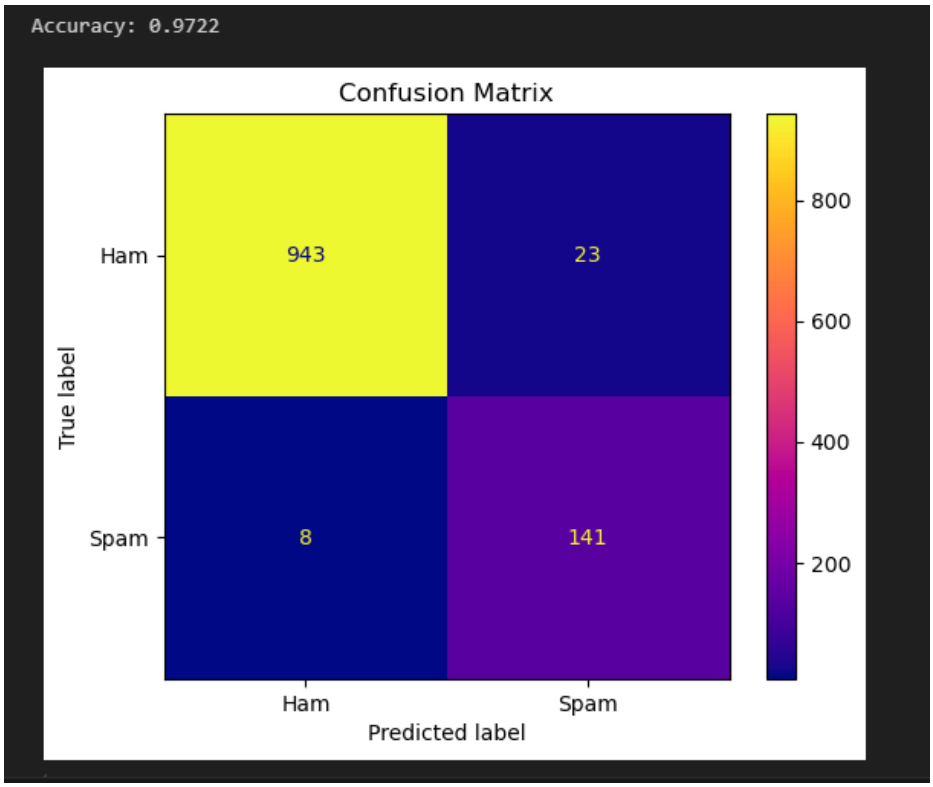
**Example**:

```
print(predict_spam("Congratulations! You've won a free iPhone. Call now!"))
```

output: “Spam”

**9. Results:**




**10. Conclusion**

This simple NLP pipeline demonstrates effective spam detection using a Multinomial Naive Bayes classifier. The accuracy and interpretability make it a great baseline for more complex models.