Text classifier Analysis Report

COURSE PRESENTER:

(Dr. Omaima Fallatah)

SUBMITTED BY:

Name	Id
امل عوض العتيبي	444001258
ريوف فيصل المحنوني	444003028
رهف ياسين برناوي	444006091

Contents

Introduction

This project aims to build a model to analyze and classify text messages into spam and ham categories using Natural Language Processing (NLP) techniques. The data was cleaned, analyzed, and a Naive Bayes model was selected to evaluate performance and determine classification accuracy.

1. Dataset

- **Dataset Name**: spam.csv
- **Columns**:
    - target: Indicates if a message is spam (1) or ham (0).
    - text: The actual text of the message.

2. Data Processing

We cleaned the dataset by:

- **Removing unnecessary columns**:

Data shape before removal:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
1480	ham	Have you always been saying welp?	NaN	NaN	NaN
2131	ham	S...from the training manual it show there is ...	NaN	NaN	NaN
2728	spam	Urgent Please call 09066612661 from landline. ...	NaN	NaN	NaN
1956	ham	K...k:)why cant you come here and search job:)	NaN	NaN	NaN
1919	ham	Yar i wanted 2 scold u yest but late already.....	NaN	NaN	NaN

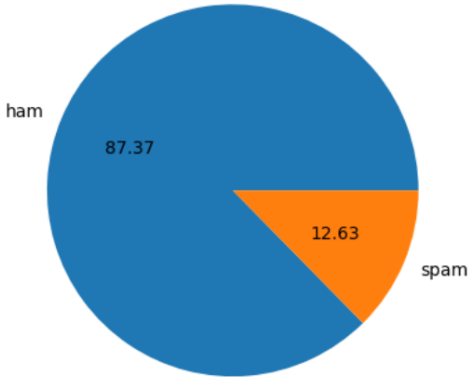
Data shape after removal:

	v1	v2
1889	ham	I gotta collect da car at 6 lei.
2258	ham	Sad story of a Man - Last week was my b'day. M...
21	ham	I□Û÷m going to try for 2 months ha ha only joking
4656	spam	PRIVATE! Your 2003 Account Statement for shows...
3165	spam	HOT LIVE FANTASIES call now 08707509020 Just 2...

- **Checking for missing values**
- **Removing duplicate values**

3. Exploratory Data Analysis (EDA)

- **Message distribution between categories**:
    - Pie chart:



    - Message ratio: (ham and spam percentage)

target	
0	4516
1	653

4. Text Preprocessing

To prepare the data for model training, we conducted several preprocessing steps to ensure the text was clean, consistent, and suitable for analysis:

1. **Standardizing Text: Converted all text to lowercase, which removed any case sensitivity in words.**
2. **Removing Symbols and Stopwords: Removed unnecessary symbols and common stopwords that don't add value to the analysis (e.g., "the," "and").**
3. **Vectorization with TF-IDF: Transformed the cleaned text into a numerical format using TF-IDF (Term Frequency-Inverse Document Frequency), focusing on the relevance of each term in relation to the whole dataset.**

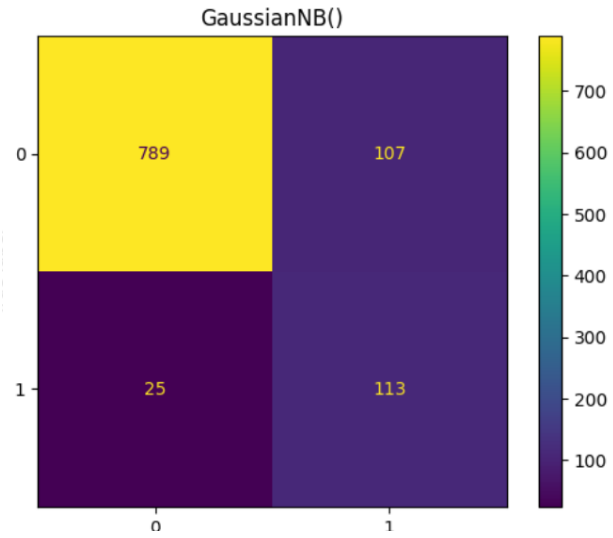
This step allowed the model to learn from word patterns effectively.

5. Model Performance

Three Naive Bayes classifiers were applied to assess model effectiveness for spam classification:

1-GaussianNB:

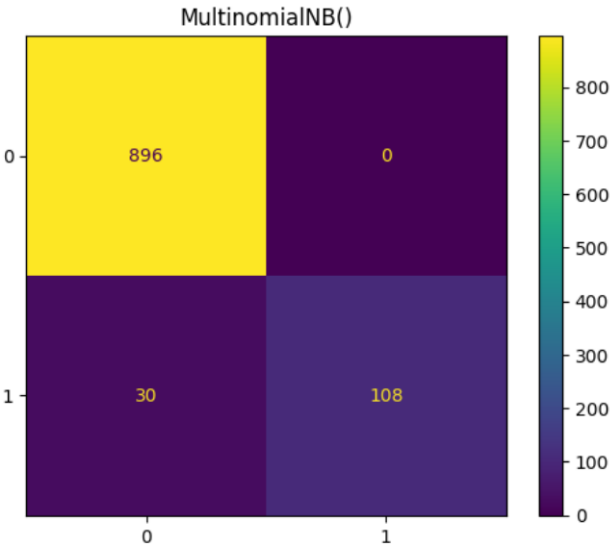
- **Accuracy : 0.87**
- **Precision : 0.51**



Lowest, as it assumes continuous data, which is less suited to discrete text features.

2-MultinomialNB:

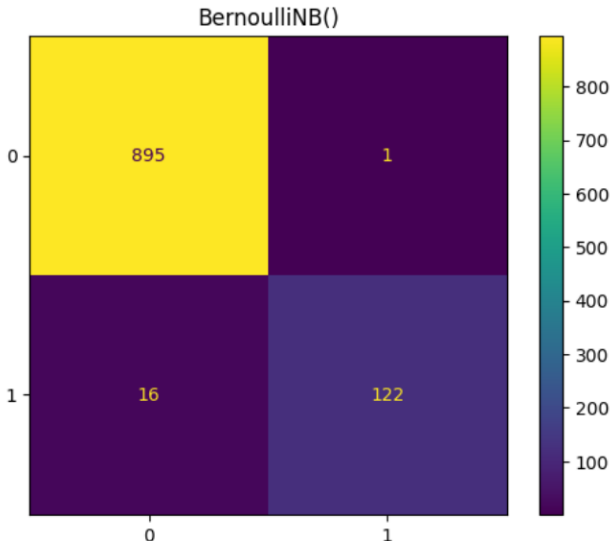
- **Accuracy : 0.97**
- **Precision : 1.0**



Performed well, especially with the frequency-based TF-IDF features.

3-BernoulliNB:

- **Accuracy : 0.98**
- **Precision : 0.99**



The highest, benefiting from binary representation, making it well-suited for spam detection.

Insights Gained from Analysis

1. **Importance of Text Preprocessing**: The TF-IDF transformation helped highlight important words tied to spam, making it easier for the models to detect patterns accurately.
2. **Best Model for Spam Detection**: The **BernoulliNB** model performed the best, as it works well with binary (yes-or-no) data, which is ideal for classifying messages as spam or not.
3. **Effective Spam Filtering**: With thorough text preprocessing, Naive Bayes models like these can achieve high accuracy, showing their strength for practical spam detection tasks.