SPAM
HAM

NATURAL LANGUAGE
PROCESSING

PROJECT-1
MAIBASSAR MIRAS

PROJECT OVERVIEW

THIS PROJECT IMPLEMENTS AN AUTOMATIC TEXT CLASSIFICATION SYSTEM THAT IDENTIFIES WHETHER A MESSAGE IS SPAM OR HAM (NON-SPAM).

THE MAIN OBJECTIVE IS TO CLASSIFY TEXT MESSAGES BASED ON THEIR CONTENT, USING NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES AND CLASSICAL MACHINE LEARNING ALGORITHMS.

THE PROJECT DEMONSTRATES THE COMPLETE NLP WORKFLOW:

- **DATA PREPROCESSING AND TEXT CLEANING**
- **MODEL TRAINING**
- **PERFORMANCE EVALUATION**

PRESENTED BY
MIRAS MAIBASSAR

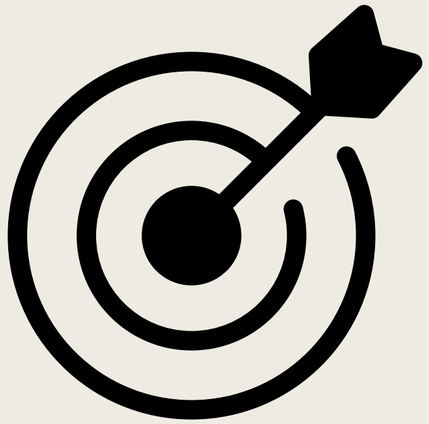
WHAT WE DO & METHODS USED

What We Do

- Load a dataset of text messages
- Clean and normalize the text data
- Convert textual labels into numerical format
- Split the dataset into:
  - training set
  - validation set
  - test set
- Train a text classification model

Methods Used

- Natural Language Processing (NLP)
- Bag-of-Words representation
- Naive Bayes classifier
- Laplace smoothing
- Log-probabilities for numerical stability



PROJECT GOAL

TO BUILD AN INTERPRETABLE AND EFFICIENT MODEL CAPABLE OF AUTOMATICALLY DETECTING SPAM MESSAGES BASED ON TEXT CONTENT.

Expected Outcome

- Accurate classification of unseen messages
- High accuracy on validation and test datasets
- Clear understanding of probabilistic models in NLP
- A scalable and extensible solution for text classification tasks

WHY THIS PROJECT MATTERS

- SPAM DETECTION IS A REAL-WORLD NLP PROBLEM
- TEXT DATA REQUIRES PREPROCESSING AND PROBABILISTIC REASONING
- NAIVE BAYES PROVIDES STRONG BASELINES FOR TEXT CLASSIFICATION
- MODEL INTERPRETABILITY IS CRITICAL FOR TRUST AND ANALYSIS

DATASET CLASS — PREPROCESSING & LABEL ENCODING

The **Dataset** class is responsible for preparing raw SMS messages
for machine learning.
It performs text cleaning, label encoding,
and prepares the data for training and evaluation.
The goal is to transform unstructured text
into a clean and numerical format suitable for a classifier.

TEXT DATA → MACHINE-READABLE FORMAT

TEXT PREPROCESSING

- LOWERCASING ALL MESSAGES
- REMOVING PUNCTUATION USING REGULAR EXPRESSIONS
- NORMALIZING RAW TEXT INPUT
THIS REDUCES NOISE AND IMPROVES MODEL PERFORMANCE.

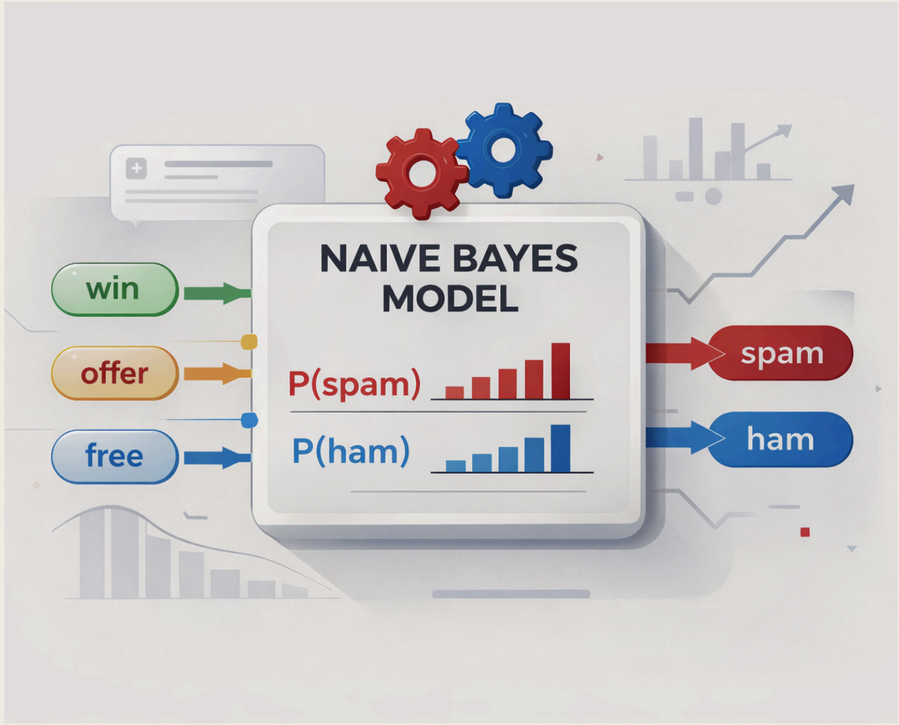
LABEL ENCODING

- CLASS LABELS ARE CONVERTED TO NUMBERS
- LABEL2NUM: LABEL → NUMERIC VALUE
- NUM2LABEL: NUMERIC VALUE → LABEL
THIS ALLOWS THE MODEL TO WORK WITH CATEGORICAL CLASSES.

PRESENTED BY
MIRAS MAIBASSAR

NAIVE BAYES MODEL

PROBABILISTIC TEXT CLASSIFICATION



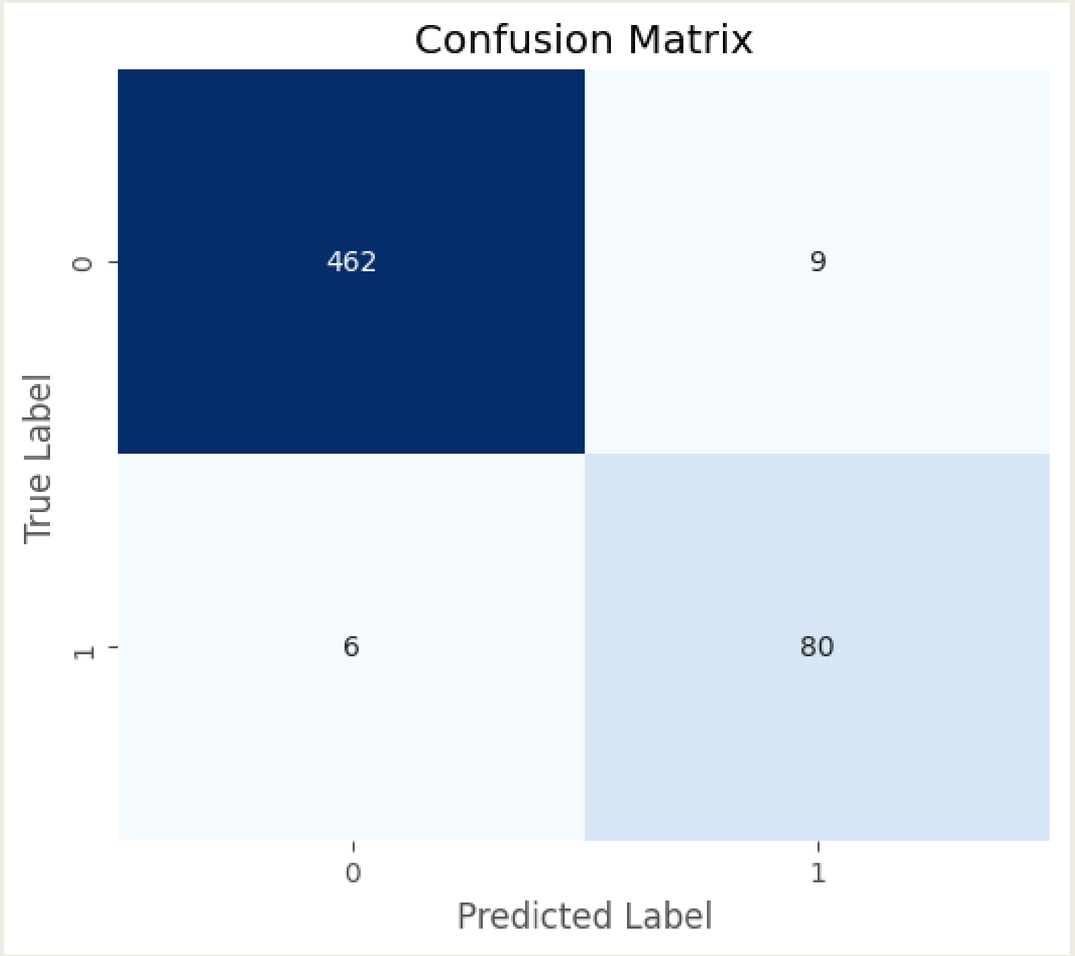
The model is a custom implementation of Multinomial Naive Bayes for SMS spam classification.
It uses word frequencies and probability theory to determine whether a message is spam or ham.
The model is trained from scratch without external machine learning libraries.

KEY CHARACTERISTICS:

- **BAG-OF-WORDS REPRESENTATION**
- **LOG-PROBABILITIES**
- **LAPLACE SMOOTHING (ALPHA)**
- **HIGH INTERPRETABILITY**

PRESENTED BY
MIRAS MAIBASSAR

CONFUSION MATRIX EXPLANATION



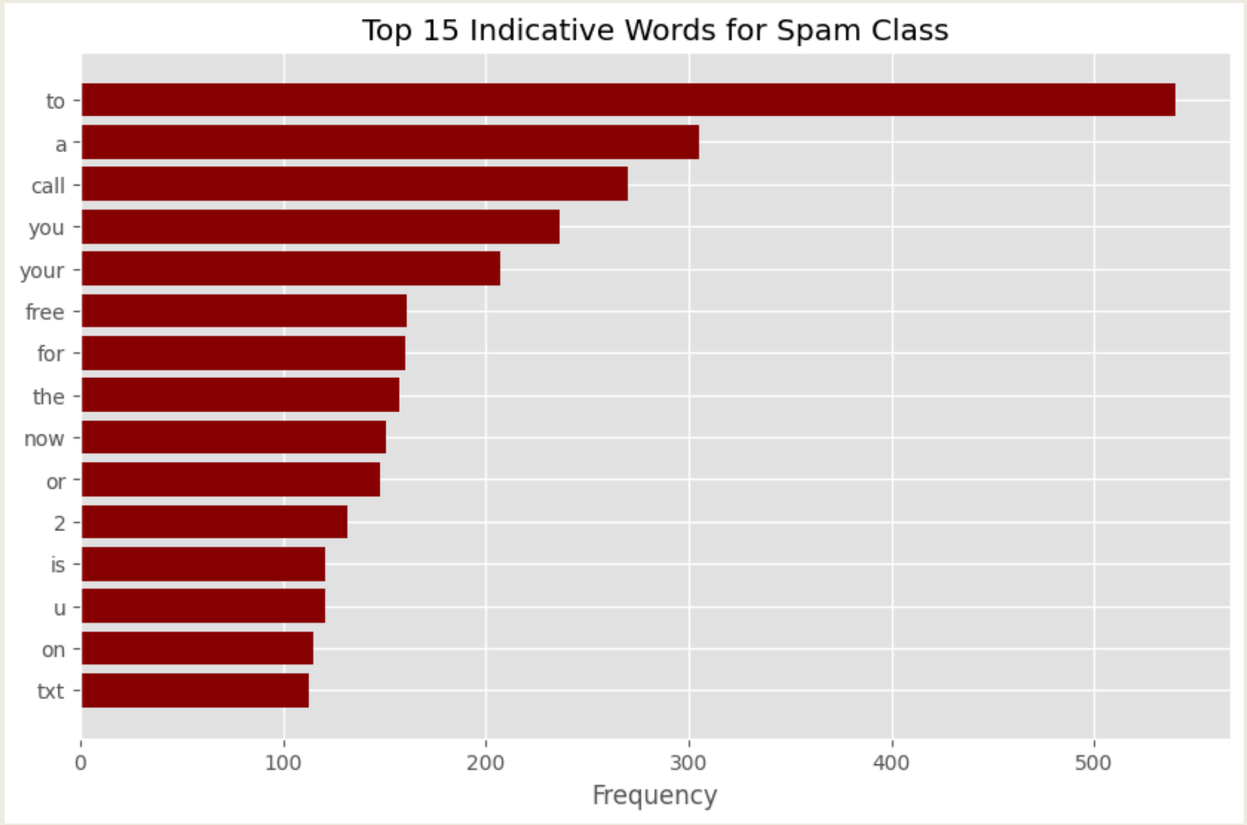
INTERPRETATION

- THE MODEL CORRECTLY CLASSIFIES THE MAJORITY OF MESSAGES.
- ONLY A SMALL NUMBER OF HAM MESSAGES ARE INCORRECTLY MARKED AS SPAM, WHICH IS IMPORTANT FOR REAL-WORLD USABILITY.
- THE MODEL DEMONSTRATES HIGH ACCURACY AND STRONG RELIABILITY, ESPECIALLY IN DETECTING SPAM MESSAGES.
- FALSE NEGATIVES (SPAM CLASSIFIED AS HAM) ARE MINIMAL, INDICATING GOOD SPAM DETECTION PERFORMANCE.

KEY INSIGHT

THIS CONFUSION MATRIX CONFIRMS THAT THE NAIVE BAYES CLASSIFIER PERFORMS EFFECTIVELY ON TEXT DATA AND ACHIEVES A GOOD BALANCE BETWEEN PRECISION AND RECALL.

Spam Word Frequency Analysis

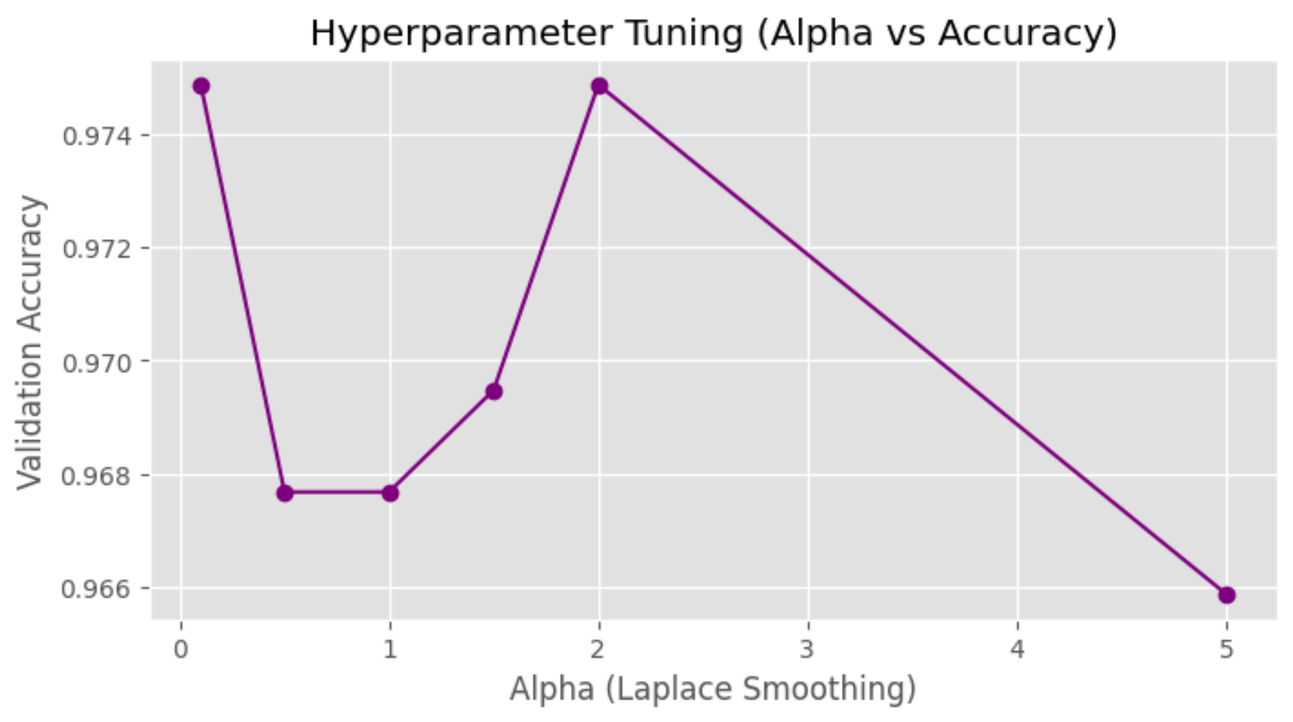


Why This Matters

- The diagram demonstrates that the model learns meaningful patterns from data, not random noise.
- It confirms the interpretability of the Naive Bayes approach: we can directly see which words influence predictions.
- Such analysis helps understand why a message is classified as spam.

TOP INDICATIVE WORDS FOR SPAM CLASS
THIS CHART SHOWS THE MOST FREQUENT WORDS APPEARING IN SPAM MESSAGES LEARNED BY THE NAIVE BAYES MODEL DURING TRAINING.
THE MODEL BUILDS A FREQUENCY-BASED REPRESENTATION OF SPAM TEXTS AND IDENTIFIES WORDS THAT ARE HIGHLY INDICATIVE OF THE SPAM CLASS.

HYPERPARAMETER TUNING (ALPHA)



Hyperparameter Optimization: Laplace Smoothing (α)

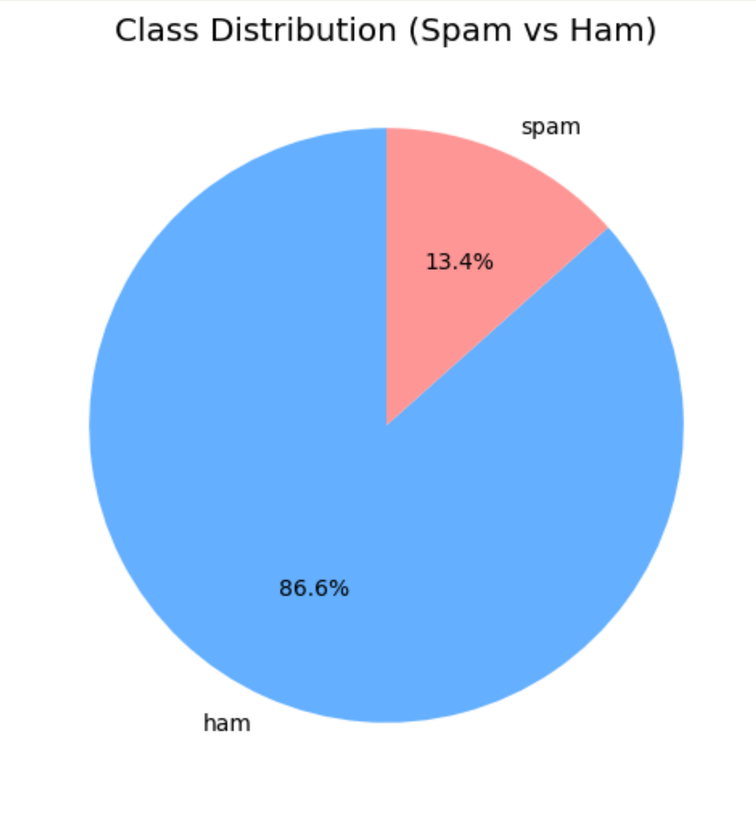
In this experiment, we performed **hyperparameter tuning** for the Naive Bayes classifier by testing different values of the **Laplace smoothing parameter (α).**

Laplace smoothing controls how the model handles rare or unseen words and helps prevent zero probabilities during classification.

RESULTS & INTERPRETATION

- **THE BEST VALIDATION PERFORMANCE WAS ACHIEVED AT A = 0.1.**
- **SMALLER VALUES OF A PROVIDED BETTER GENERALIZATION FOR THIS DATASET.**
- **LARGER A VALUES LED TO EXCESSIVE SMOOTHING, SLIGHTLY REDUCING ACCURACY.**

CLASS DISTRIBUTION (EDA)



CLASS DISTRIBUTION: SPAM VS HAM

THIS CHART ILLUSTRATES THE DISTRIBUTION OF CLASSES IN THE DATASET USED FOR TRAINING AND EVALUATION.

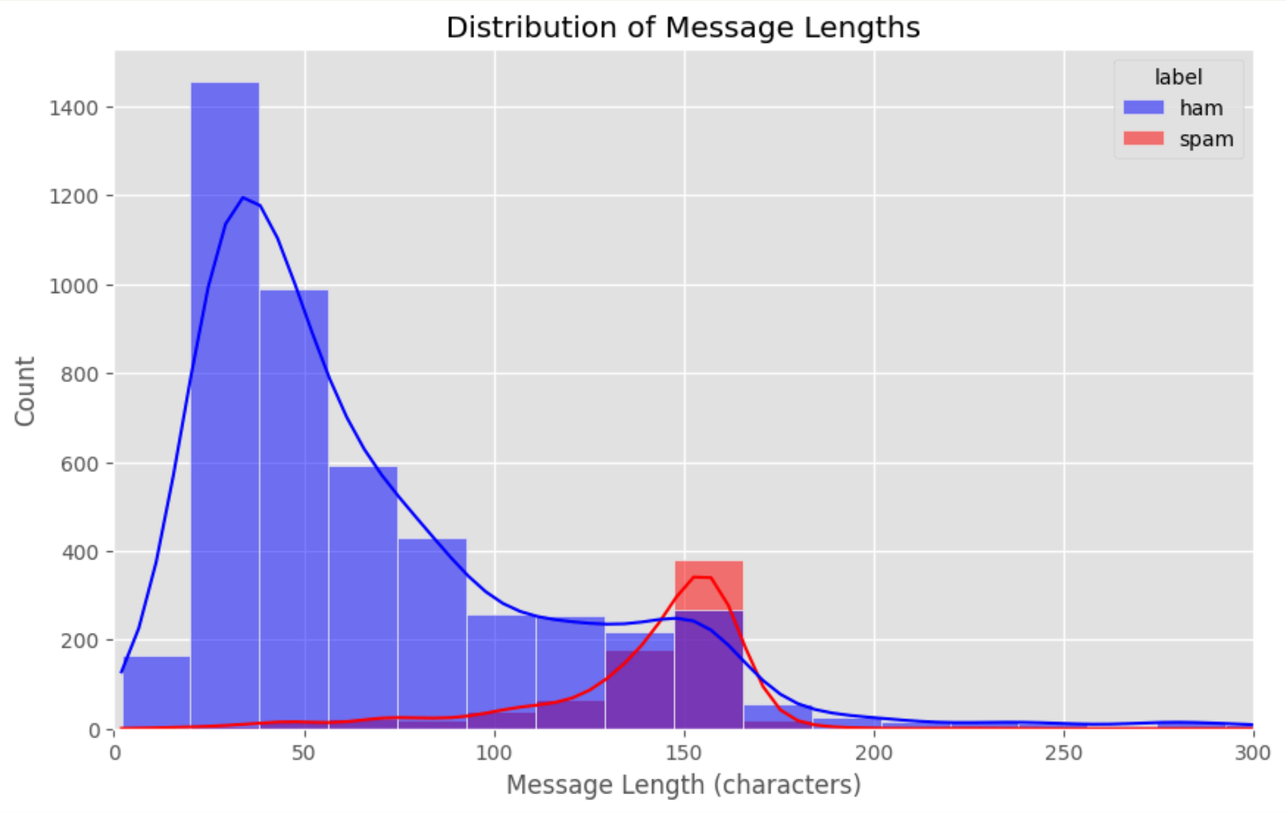
- Total samples: 5,572
- Ham messages: 4,825 (≈ 86.6%)
- Spam messages: 747 (≈ 13.4%)

KEY OBSERVATIONS

- The dataset is imbalanced, with a significantly larger number of ham messages.
- Spam messages represent a smaller but important portion of the data.
- Such imbalance is typical for real-world spam detection tasks.

MESSAGE LENGTH ANALYSIS (EDA)

Distribution of Message Lengths
This plot shows the distribution of message lengths (in characters) for both classes: ham and spam.



Key Observations

- Ham messages are generally shorter and concentrated at lower character counts.
- Spam messages tend to be longer on average, often containing promotional or informational content.
- There is a noticeable shift in the distribution of spam messages toward higher message lengths.

WHY THIS MATTERS

- MESSAGE LENGTH PROVIDES AN ADDITIONAL DISCRIMINATIVE SIGNAL FOR CLASSIFICATION.
- LONGER MESSAGES OFTEN INCLUDE:
  - ADVERTISEMENTS
  - CALL-TO-ACTION PHRASES
  - DETAILED OFFERS
- THIS SUPPORTS THE EFFECTIVENESS OF WORD-BASED PROBABILISTIC MODELS SUCH AS NAIVE BAYES.

MODEL INFERENCE EXAMPLES

Real-World Message Classification
This slide demonstrates how the trained Naive Bayes model classifies previously unseen messages.
Several real-world–like examples were passed to the model to evaluate its practical behavior.

Example Predictions

- **“You have been selected to receive a £1000 cash prize…”**
- **→ Spam**
- **“I’ll be at the office around 10 am. See you then.”**
- **→ Ham**
- **“Your Apple ID has been locked due to unauthorized login attempts…”**
- **→ Spam**
- **“Did you finish the report for tomorrow?”**
- **→ Ham**
- **“Shop now! 50% off on all items…”**
- **→ Spam**

Interpretation

- The model correctly identifies promotional, urgent, and phishing-like messages as spam.
- Neutral and conversational messages are classified as ham.
- This confirms that the model generalizes well beyond the training dataset.

Key Insight
The inference results show that the classifier is practically usable, interpretable, and effective for real-world spam detection tasks.