1. <u>Email Spam Classification</u>
2. 1. Introduction

The purpose of this project is to build a machine learning model that classifies email messages as either **Spam** or **Ham** (not spam). Spam detection is an important real-world application of Natural Language Processing (NLP) used to filter unwanted or harmful messages.

3. 2. Dataset

- Dataset: `spam.csv`
- Total Records: ~5,500 messages
- Columns Used:
    - **Category**: Label (`ham` or `spam`)
    - **Message**: The text content of the email/SMS message

4. 3. Methodology

The project was completed in three main phases:

5. Data Cleaning & Preprocessing

- Converted all text to lowercase.
- Removed punctuation, numbers, and special characters.
- Tokenized messages into words.
- Removed stopwords using NLTK.
- Applied stemming with Porter Stemmer.
- Joined tokens back into cleaned sentences.

6. Vectorization

- Used **TF-IDF Vectorizer** with n-grams to convert text into numerical form.
- Parameters:
    - `lowercase=True`
    - `ngram_range=(1,2)`
    - `min_df=1` (to keep rare but important words)

7.   Model Training & Evaluation

- Models Used:
    - Multinomial Naive Bayes
    - Logistic Regression
    - Linear Support Vector Machine (SVM)
- Evaluation Metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-Score
    - Confusion Matrix

8.   Visualizations

- **Class Distribution**: Showed number of spam vs ham messages in the dataset.
- **Confusion Matrix Heatmap**: Displayed model performance visually.

9.   Results

Model	Accuracy	Precision (Spam)	Recall (Spam)	F1-Score (Spam)
Multinomial Naive Bayes	0.971	0.962	0.905	0.933
Logistic Regression	0.982	0.977	0.934	0.955
Linear SVM (Best Model)	0.984	0.981	0.934	0.957

Best Model: Linear SVM – chosen based on overall highest F1-Score.

10.   Predictions on Sample Messages

The model was tested with unseen messages. Examples:

| Message	Prediction
Congratulations! You have won a $1000 Walmart gift card. Click here...	Spam
Are we still meeting at 3pm today?	Ham
URGENT: Your account is locked. Verify your details immediately.	Spam
Ok, see you soon!	Ham
WINNER! Free vacation to Bahamas...	Spam

11.   Conclusion

The spam classification system successfully distinguishes spam and ham messages with high accuracy. Using TF-IDF with n-grams and the Linear SVM classifier provided the best performance. Future improvements can include:

- Adding more training data.
- Trying deep learning models like LSTM or BERT.
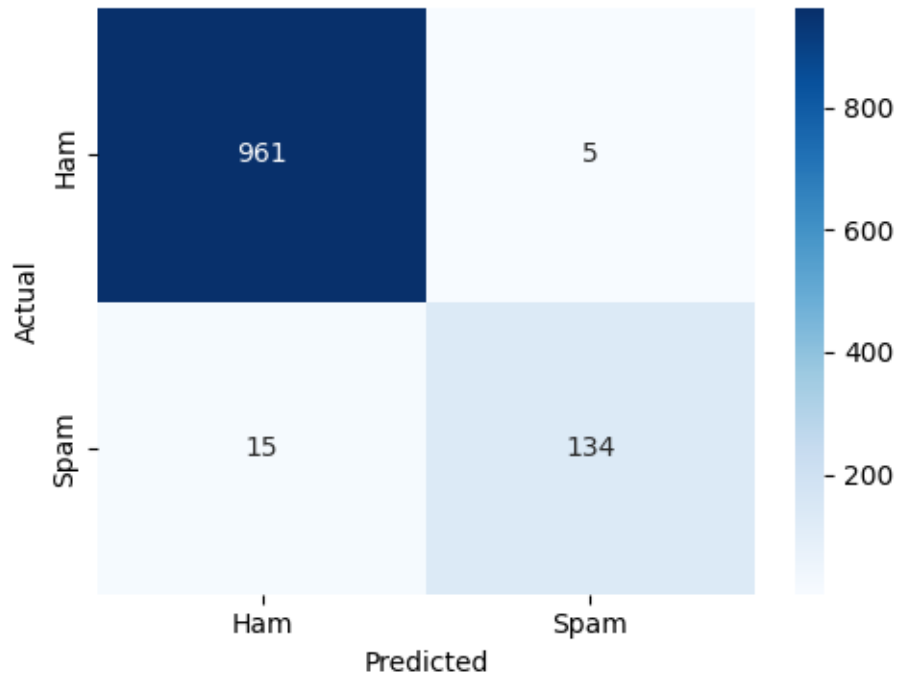- Using lemmatization instead of stemming.

Spam vs Ham count
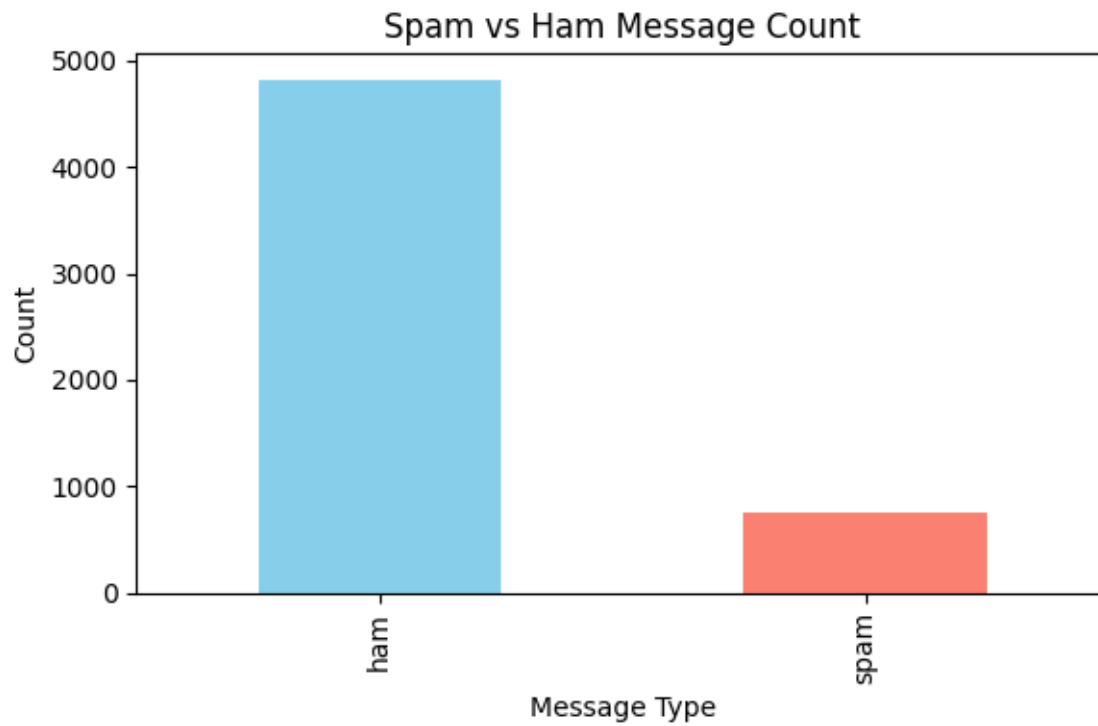
Confusion Matrix for best model

Confusion Matrix - Best Model (Linear SVM)