SMS Spam Detection-06/10/25

SMS Spam Detection Project: Final Report

1. Project Objective

The primary goal of this project was to build and evaluate a machine learning model capable of accurately classifying SMS messages as either "ham" (legitimate) or "spam" (unsolicited commercial/malicious). This task served as an introduction to fundamental Natural Language Processing (NLP) techniques and classification algorithms.

2. Data Preparation and Text Preprocessing

The project began with the **SMS Spam Collection Dataset** (5,572 total messages).

A. Initial Cleaning

- **Duplicate and Missing Values:** Duplicate messages (403 rows removed) were handled, and missing values were dropped.
- **Label Encoding:** Text labels were converted to numerical form: **Ham →0** and **Spam →1**.

B. Text Preprocessing Steps

The raw message text was transformed into a standardized format for machine learning:

1. **Normalization:** Converted all text to **lowercase**.
2. **Removal:** Eliminated punctuation, numbers, URLs, and HTML tags.
3. **Tokenization:** Broke messages into individual words (tokens).
4. **Stopword Removal:** Eliminated common, low-information words (e.g., 'the', 'a', 'is').
5. **Lemmatization:** Reduced words to their base or root form (e.g., 'running' → 'run').

---

3. Feature Extraction

Text data was converted into numerical vectors for machine learning using **TF-IDF**.

- **Method: TF-IDF (Term Frequency-Inverse Document Frequency)** was used to assign weights to words. TF-IDF gives a higher score to words that are **frequent in a specific message** (high Term Frequency) but **rare across the entire dataset** (high Inverse Document Frequency), effectively highlighting important, unique words (like 'URGENT' or 'PRIZE') that indicate spam.
- **Data Split:** The dataset was split into training and testing sets (70% train, 30% test) using a **stratified approach** to ensure the proportion of spam (≈13.4%) was maintained in both subsets.

---

4. Model Training and Evaluation

Two different classification algorithms were trained on the TF-IDF feature vectors and evaluated using the test set (1,551 messages).

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9716	0.9691	0.801	0.8771
Naive Bayes (Multinomial)	0.9587	1	0.6735	0.8049

A. Model Comparison Table

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9716	0.9691	0.801	0.8771
Naive Bayes (Multinomial)	0.9587	1	0.6735	0.8049

B. Key Mathematical Metrics

Accuracy=TotalTP+TNPrecision=TP+FPTPRecall=TP+FNTP

F1-Score=2×[(Precision+Recall) / (Precision×Recall)]

C. Final Model Selection

The **Logistic Regression** model was selected as the final solution:

- It achieved the **highest F1-score (0.8771)**, representing the best balance between catching spam and avoiding false alarms.
- It provided significantly better **Recall (80.10% vs. 67.35% for NB)**, meaning it missed fewer actual spam messages.
- **5. Final Model Performance Analysis (Logistic Regression)**
- The confusion matrix for the final Logistic Regression model showed the following results on the test set:

Outcome	Count	Interpretation
True Positives (TP)	157	Correctly identified Spam.
True Negatives (TN)	1350	Correctly identified Ham.
False Positives (FP)	5	Ham incorrectly marked as Spam (Legitimate messages lost).
False Negatives (FN)	39	Spam incorrectly marked as Ham (Spam leaked to the inbox).

The high **Precision** (96.91%) and very low **False Positive** count (5) are ideal for a user-facing spam filter, as user experience dictates that legitimate messages must not be blocked.

6. Deliverables and Conclusion

The project successfully generated all required deliverables:

- A **Cleaned Dataset** (df['clean_message']).
- Two **Trained Models** (Naive Bayes and Logistic Regression).
- Complete **Evaluation Metrics**.
- **Visualizations** (Confusion Matrix and class distribution).
- A **Reusable Script** with functions to predict new messages.
- **Saved Models** (Logistic Regression model and TF-IDF vectorizer) for future use.

The final **Logistic Regression model** provides robust and reliable spam detection, achieving high accuracy while effectively prioritizing the safety of legitimate user messages.