Full Report: Email Spam Classifier

Name: Saima Chowdhury
Email: saima.chowdhury811@gmail.com

1. Collecting Dataset

Dataset Name: SMS Spam Collection Dataset (Kaggle)
Dataset Link: *https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset/suggestions*
Description: Contains 5,574 messages in English
Columns: label (spam/ham), message (text)
Target: Label.

*Renamed necessary columns according to requirements (v1' as 'label(spam/harm)', 'v2' as 'message(text)).
*Dropped unnecessary columns (Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4)

2. Cleaning and Preprocessing

Steps applied to raw text:

1) Text normalization (lowercasing)
2) Remove punctuation, special characters, numbers
3) Tokenization & stop-word removal
4) Vectorization: Bag of Words (CountVectorizer), TF-IDF (TfidfVectorizer)

3. Building and Training Models

Comparison of given three models-

- Naive Bayes:
  Pro: Very fast and works surprisingly well with sparse text features.
  Con: Assumes independence between words (not realistic).
- Logistic Regression:
  Pro: Usually more accurate than NB with TF-IDF and gives interpretable weights.
  Con: Needs more computation and careful regularization to avoid overfitting.
- Random Forest:
  Pro: Can model complex non-linear relationships in data with dense, engineered features (e.g., embeddings, topic vectors).
  Con: Performs poorly on high-dimensional sparse text (like bag-of-words).

After comparing the given three models, considering the advantages and disadvantages, Naïve Bayes seemed the most appropriate model for sparse data/text classification problems.

4. Evaluation Metrics

Dataset split: 70% train, 30% test

Metrics used:

- Accuracy: ~94.8%
- Confusion Matrix:

Model Name	Accuracy	Confusion Matrix			
		TP	**TN**	**FP**	**FN**
Naïve Bayes	~94.8%	132	1453	0	87
Logistic Regression	~95.1%	138	1452	1	81
Random Forest	~97.1%	171	1453	0	48

5. Comparative analysis of algorithms (e.g., Naive Bayes vs Logistic Regression vs Random Forest)

◉ Accuracy Comparison with Bar-plots:

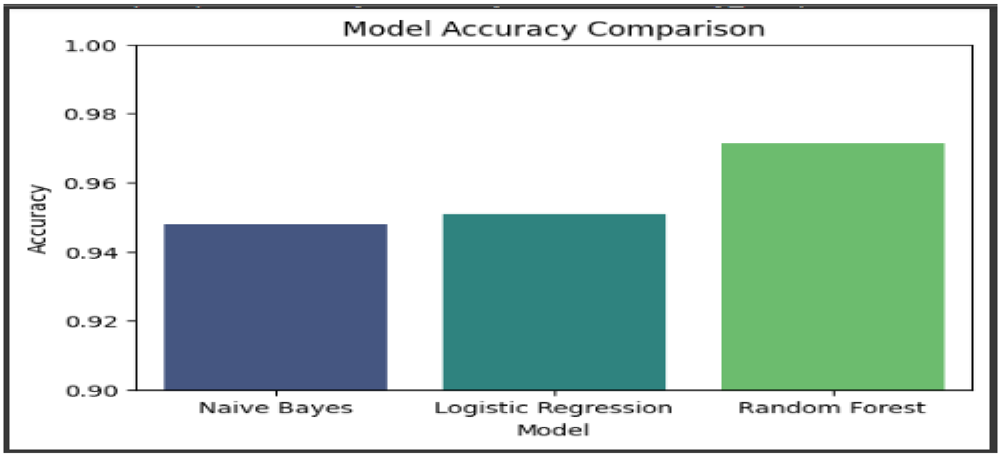

Fig.01: Accuracy Comparison of NB, LR, RF Models.

◉ Confusion Matrix Comparison with Heat-maps:

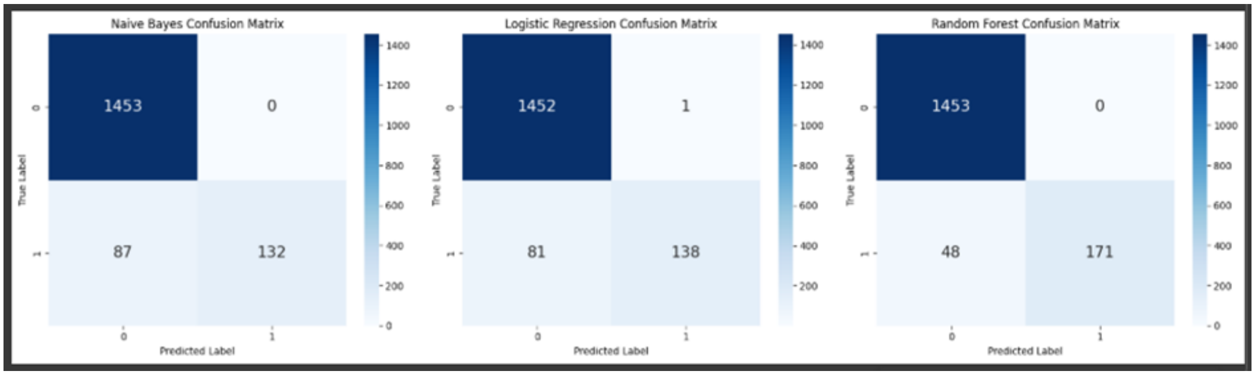

Fig.02: Confusion Matrix Comparison of NB, LR, RF Models.

On looking at the comparative analysis, the Random Forest model is the best performing model for this email spam classification task using TF-IDF features. It demonstrates the highest overall accuracy and is most effective at correctly identifying 'spam' messages while avoiding misclassifying 'ham' messages.

4. Conclusion

Based on these metrics the breakdowns:

- All three evaluated models demonstrated high accuracy in classifying spam and ham messages, and provided the right prediction during verification with input data, indicating their suitability for this task.

- While **Naive Bayes Model** called a strong baseline for text classification and performs surprisingly well with sparse data. In this case, it also provided an advisable accuracy. Apart from that, the **Logistic Regression Model** also provided very high accuracy. Although **Random Forest Model** performs poorly on high-dimensional sparse text (like bag-of-words) but, **for this particular dataset and with TF-IDF features**, it yielded better results here.

-End-