Spam Email Filter: Report

Steps:

1- **Data cleaning and preprocessing:**

- **Remove Email Addresses , URLs and Remove Non-Alphanumeric Characters and Punctuation:** Removed email addresses and URLs, as these elements do not provide meaningful information for spam detection.
- **Tokenization and Lowercasing:** The text was tokenized into words and converted to lowercase. This helps to standardize the text and makes it easier to identify relevant features then be tokenized
- **Removing Punctuation, Numbers, and Stopwords:** Punctuation and numbers were removed as they do not contribute to the semantic meaning of the text. Common stopwords were removed to focus on significant terms.
- **Stemming:** Stemming was applied to reduce words to their base form, helps to generalize the features.

2- **Data splitting**:

we choosed the portion (60-40), 40 for testing and 60 for training. to split the data randomly to avoid bias.

3- **Model Training and Text Embedding**:

- ○ The **classifiers** chosen for training the data are Logistic Regression and Decision Tree Classifier. These classifiers were selected due to their simplicity, interpretability, and effectiveness in handling both linear and non-linear relationships in the data.
- ○ **For text embedding techniques, we choosed those for the following reasons:**
  - **Doc2Vec**: It captures the semantic meaning of the text along with its contextual information, making it suitable for document-level embeddings.
  - **TF-IDF**: Represents importance of each word in the document relative to the entire corpus, which helps in capturing unique features of each document.
  - **Bag of Words**: Its simplicity and effectiveness in capturing the occurrence of words in the document without considering their order.
  - **Word2Vec**: Utilized to generate word embeddings based on the semantic meanings of words, which can capture the context and relationships between words.

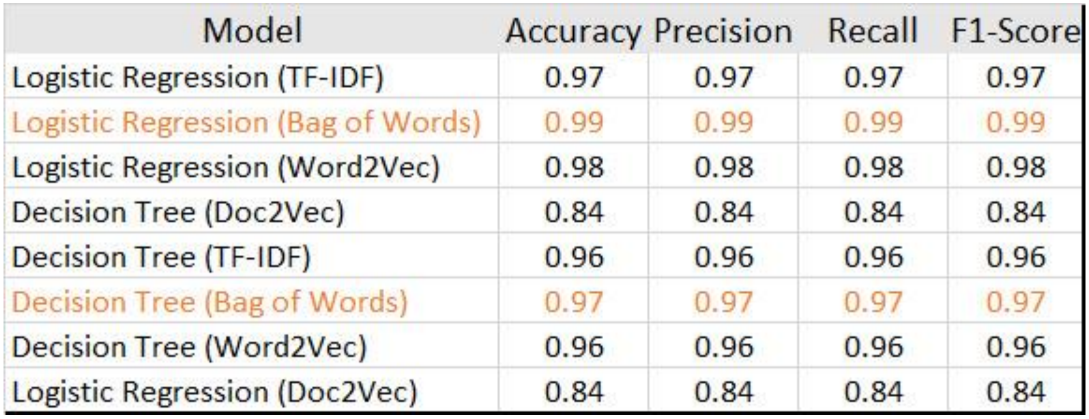
The models were evaluated using accuracy, precision, recall, and f1-score metrics. These metrics provide a comprehensive evaluation of the model's performance.

(The green parts are based on neural networks)

- **Path1:**
  - ○ **Classifier:** Logistic Regression
  - ○ **Text Embedding Technique:** Bag of Words
- **Path 2:**
  - ○ **Classifier:** Logistic Regression
  - ○ **Text Embedding Technique:** TF-IDF
- **Path 3:**
  - ○ **Classifier:** Logistic Regression
  - ○ **Text Embedding Technique:** Word2Vec (Based on neural networks)
- **Path 4:**
  - ○ **Classifier:** Logistic Regression
  - ○ **Text Embedding Technique:** Doc2Vec (Based on neural networks)
- **Path 5:**
  - ○ **Classifier:** Decision Tree

    - **Text Embedding Technique:** Bag of Words
- **Path 6:**
    - **Classifier:** Decision Tree
    - **Text Embedding Technique:** TF-IDF
- **Path 7:**
    - **Classifier:** Decision Tree
    - **Text Embedding Technique:** Word2Vec (Based on neural networks)
- **Path 8:**
    - **Classifier:** Decision Tree
    - **Text Embedding Technique:** Doc2Vec (Based on neural networks)

4- Results and Dominant Models:



Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression (TF-IDF)	0.97	0.97	0.97	0.97
Logistic Regression (Bag of Words)	0.99	0.99	0.99	0.99
Logistic Regression (Word2Vec)	0.98	0.98	0.98	0.98
Decision Tree (Doc2Vec)	0.84	0.84	0.84	0.84
Decision Tree (TF-IDF)	0.96	0.96	0.96	0.96
Decision Tree (Bag of Words)	0.97	0.97	0.97	0.97
Decision Tree (Word2Vec)	0.96	0.96	0.96	0.96
Logistic Regression (Doc2Vec)	0.84	0.84	0.84	0.84

Based on the provided table, the two dominant models are:

1. Logistic Regression (Bag of Words)
2. Decision Tree (Bag of Words)

These models were chosen as dominant models because they achieved the highest accuracy, precision, recall, and F1-score among all tested models.

Reasons for Dominance:

1. **Logistic Regression (Bag of Words):**
    - This model demonstrated excellent performance across all evaluation metrics, achieving the highest possible values.
    - Logistic Regression is effective for binary classification tasks like spam detection.
    - The Bag of Words (BoW) feature extraction technique is simple yet effective in representing text data, leading to a highly accurate model.
2. **Decision Tree (Bag of Words):**
    - This model also achieved high performance across all evaluation metrics.
    - Decision Trees are interpretable and capable of handling both numerical and categorical data.
    - The BoW feature extraction technique worked well with the Decision Tree classifier, resulting in a highly effective spam detection model.