# **Report: Email Spam Classification Project**

## Notebook Flow Description

The notebook follows a structured process flow for email spam classification:

1. **Data Loading**: Importing necessary libraries and loading the dataset.
2. **Data Preprocessing**: Cleaning the email text data by removing headers, HTML tags, URLs, and non-alphabetic characters. Tokenization, stopword removal, and stemming are applied.
3. **Data Splitting**: Dividing the dataset into training and testing sets for model evaluation.
4. **Text Embedding**: Generating Word2Vec and Doc2Vec embeddings for text representation.
5. **Model Training**: Training decision tree and logistic regression classifiers using both Word2Vec and Doc2Vec embeddings.
6. **Model Evaluation**: Evaluating trained models using accuracy and precision metrics.

The order of operations is chosen to ensure proper data preprocessing before feature extraction and model training. For example, stemming is applied after tokenization to ensure individual word tokens are properly stemmed.

## **Data Preprocessing & Features Extraction**

* **Text Cleansing**: Removing email headers, HTML tags, URLs, and non-alphabetic characters to ensure clean text data.
* **Tokenization**: Breaking down text into individual word tokens for further processing.
* **Stop word Removal**: Eliminating common stop words to focus on meaningful content.
* **Stemming & Lemmatization**: Reducing words to their base forms to improve model generalization.

These techniques are chosen to address common issues in text data such as noise, redundancy, and variation in word forms.

## **Data Splitting**

The dataset is split into training and testing sets with a 60:40 ratio. This ratio is chosen to allocate a sufficient amount of data for model training while retaining a sizable portion for evaluation.

## **Model Training**

Decision tree and logistic regression classifiers are chosen for model training due to their simplicity, interpretability, and effectiveness in text classification tasks. Word2Vec and Doc2Vec embeddings are utilized for text representation to capture semantic relationships between words and documents.

## **Model Evaluation**

Test accuracy and precision metrics are chosen to evaluate model performance. Accuracy measures the overall correctness of predictions, while precision focuses on the accuracy of positive predictions (spam emails). These metrics provide insights into both overall performance and the model's ability to correctly identify spam emails.

## **Dominant Models**

|  |  |  |
| --- | --- | --- |
| Model | Test Accuracy | Test Precision |
| Logistic Regression (Word2Vec) | 98.50% | 98.70% |
| Logistic Regression (Doc2Vec) | 96.0% | 94.50% |
| Decision Tree (Word2Vec) | 97.90% | 98.0% |
| Decision Tree (Doc2Vec) | 85.50% | 76.70% |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

The choice of Logistic Regression (Word2Vec) and Decision Tree (Word2Vec) as the dominant models is based on their superior performance in terms of accuracy compared to the other models evaluated in the project.

Both models leverage Word2Vec embeddings, which capture semantic relationships between words in the text data. This embedding technique proves to be effective in representing text features for classification tasks, contributing to the success of both logistic regression and decision tree models. Therefore, considering their strong accuracy performance and interpretability, Logistic Regression (Word2Vec) and Decision Tree (Word2Vec) are chosen as the dominant models for email spam classification.