Assignment 2 Report

Notebook Flow Description:

We followed a process of pre-processing by first data-cleansing then tokenization and lemmatization, after that we used 4 different text embedding techniques and then evaluated each with 2 different model evaluation techniques to achieve best possible outcome.

Data Preprocessing & Features Extraction:

The data only has 1 feature essentially (the email), but in a more specific manner it has numerous if you view them as individual tokens of the email itself, first we remove the html tags using beautifulsoup, then we run a regex to remove words with special characters or numbers and convert the remaining to lower case before tokenizing and lemmatizing them.

Data Splitting:

In almost all the word embedding techniques we use the full data set while splitting it by a 50/50 ratio to reduce time taken to train the model.

The exception is the Bert word embedding technique (Bidirectional Encoder Representations from Transformers) as it is very computationally expensive we had to reduce the size of the dataset to 5% of it's original size and even then it still takes around 25 minutes to execute word embedding of the whole subset.

Model Training:

Since it is a classification we chose the most efficient classification technique Logistic Regression, along side with KNN (K-Nearest Neighbors) and Decision Tree to find alternative results, we also used 2 non Neural Network based word embedding techniques N-grams and Term Frequency Inverse Document Frequency, and 2 Neural Network based word embedding techniques Word to Vector and BERT (Bidirectional Encoder Representations from Transformers) to get the most inclusive results from our models.

Model Evaluation:

We used Accuracy, Recall and precision as they are the most commonly used metrics of evaluation.

Dominant Models:

label	accuracy
tfidfKNN	0.959282
tfidfLogisticRegression	0.978951
ngramKNN	0.948930
ngramLogisticRegression	0.993789
W2VKNN	0.908903
W2VDecisionTree	0.854382
bertLogisticRegression	0.910345
bertDecisionTree	0.737931

Term Frequency-Inverse Document Frequency with Logistic Regression classification scored almost 98% accuracy but it fell second to the exceptional performance of N-gram technique (unigram) with Logistic Regression classification as it scored almost 99.4% accuracy, with that being said those are the two most dominant models.