NLP Assignment 2 Spam E-mail Filter

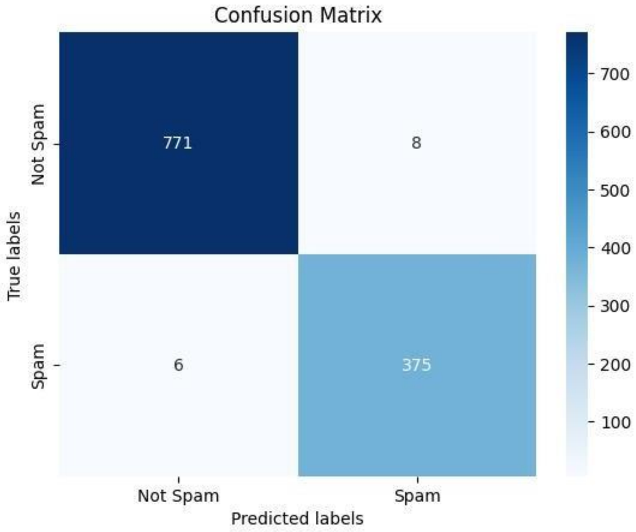
Abubaker Saeed Omer__20200798
Tonaja Mohamed__20200900
Ahmed Esaam__20200853

The code link

The notebook flow:

- First I imported some of the needed libraries, read the dataset and separate the data as in the csv file. Then I split the data randomly (80-20) it's better in most cases and helps ensure that the two sets have a similar distribution of data and that’s gives a higher accuracy and prevent overfitting (I've tried (50-50) and it was bad)
- Then preprocessing the data. This part is to make sure that we only use the important part in each email which is the text,
  1. Remove the hyperlinks
  2. Remove the numbers
  3. Remove the punctuation
  4. Remove the spaces
  5. Replace the new line
- Then I had an issue that the model considered part of the labels were float, so I used LabelEncoder() that encode target labels with value between 0 and n_classes-1.
- Then the tokenization and padding part, I set the size of the word vector as 100, number of rows in the embedding vector as 50000 and max length 2000

The first model is a **Nueral Network**, And I chose it because it has the ability to learn complex patterns and extract features from raw data and can learn to recognize important textual patterns, such as specific words, phrases, or combinations of words that are indicative of spam content.



The second model is **Logistic regression** and I chose it because it's simple and allows you to identify the importance of different features in determining whether an email is spam or not. By examining the magnitude of the coefficients, you can gain insights into which words, phrases, or other features have a significant impact on the classification decision.

- **Model Evaluation**

The Model	Highest Accuracy	Recall
Nueral Network	After the 10th epoch: 0.9985	0.9843
Logistic regression	0.9982	0.9972

Note:
More details in the code notebook