Report on Email Click Prediction

---

1) Dataset Preprocessing -

a) Remove special characters/ non alphanumeric words
b) Convert the words to lowercase
c) Remove stopwords
d) Convert the target to numeric values using encoding

2) Some important data statistics obtained are given below -

a) Word count in the body
b) Average word count per subject
c) Length of sentences
d) Target distribution

3) Models Development -

a) Feature engineering : We first build a dataframe then build a word index mapping using dictionaries. Finally the data is modified as vectors to be passed to ML models
b) Model selection : SVM for classification and Multiclass Neural Network were used because they deal with multiclass outputs. The target had four classes {0,1,2,3}.
c) Training : I was not able to perform a lot of hyperparameter tuning. The main modification is the batch size and number of epochs used for the Neural Network model

4) Model Evaluation -

Accuracy score was the main model evaluation since the target was categorical. SVM had an accuracy of 47 and Neural network had an accuracy of 37. In SVM 30% of the data was used to test the accuracy but in NN only the last batch was used.

5) Future scope -

RNN can be implemented where the input is a sequence and the output is passed to a softmax. Additionally embedding values can be used instead of directly passing the words to Neural Network models. I was unable to implement an RNN using GloVe embeddings in time.