

NLP Assignment 3: News Topic Classification; Task 2

--Utsav Basu (20CS30057)

Problem Statement

You are given a collection of news articles. Your task is to classify the news texts into 4 topics/classes—World, Sports, Business, and Science/Technology. You will first use the Word2Vec and RNN model for Task1 and then improve it by using BERT model in Task2.

Dataset

The dataset is as follows:

- It contains a train and a test dataset, each consisting of multiple sentences (Column1) and their respective labels (Column2). The labels are:
 - World (0)
 - Sports (1)
 - Business (2)
 - Sci/Tech (3)
- The train dataset contains 2000 sentences, while the test dataset contains 500 sentences. The train dataset is equally distributed, i.e., it has 25% samples from each of the four classes.
- Consider 10% randomly selected samples from the training set as a validation set

Methods

“Roberta-base” for sequence classification was used for the task. The tokenizer was used to produce the word embedding for each sentence which was then passed on to the Bert model for training/testing.

For each model the hyperparameters are as follows

- Output dims: 4 (no. of classes)
- Batch size: 32
- Sequence length: 262 (max length of tokens in train set)
- Learning rate: 10^{-5}
- Optimizer user: AdamW
- LR Scheduler used: ReduceLROnPlateau(mode='min', factor=0.5, patience=20)
- Epochs: 10

Results

The results are as follows. The Classification Score Tables and Confusion Matrices are in the notebooks:

Accuracy: 88.8000%

F1: 88.7141%