NLP Assignment Report

##### ****Objective:****

##### The goal of this assignment is to build a text classification model using the Hugging Face library to classify a dataset of text into one of multiple categories. The candidate will use a pre-trained model such as BERT or GPT-2 as a starting point and fine-tune it on the classification task.

##### ****Dataset Specifications:****

* There are two CSV Files which are - training and test CSVs which includes three columns each which are: Class Index, Title, Description.
* This is a classification problem where the dataset contains three distinct classes: World, Sports, Business, Science/Tech.
* The Class Index represents the category of the article where 1 denotes World, 2 for Sports, 3 for Business, and 4 for Science/Tech.
* The “Title” column contains the title of the articles.
* The “Description” column contains the main body content of the article.
* Source to Data-Set used- [***AG News Classification Dataset***](https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset)

##### ****Pre-Processing:****

* The pre-processing steps taken – Cleaning of Dataset
* We Merge the two diff columns present in our dataset that is title and Description.
* Drop any row with Null values.
* We Re-Assign the Labels to start with 0 rather than 1. (Just by subtracting 1 from the Class Index column)
* Then cleaning the Data using Regular expressions to Remove special Characters and Unnecessary Words and finally converting all the letters to Lower case.
* Finally, we check for the whitespaces if any.

##### ****Summary:****

* Use a pre-trained TensorFlow Hugging Face Transformer to fine-tune the model with the obtained data.
* Use the Evaluation Metrics like Confusion Matrix and determine how good the model really is at prediction.