Assignment Report

The goal is to build a text classification model using the Hugging Face library to classify a dataset text into multiple categories. The 'Go Emotions PT-BR' dataset from Kaggle is used for the training. The final model will try to identify triggered emotions out of the available 29.

Dataset link: https://www.kaggle.com/datasets/antoniomenezes/go-emotions-ptbr

Preprocessing:-

At first, all the punctuations from the text were filtered out using 'string.punctuation'. Compete for text was converted into lower case, then the 'WordNetLemmatizer' from NLTK lemmatized it.

Architecture:-

The 'DistilBert' architecture(transformer) available on Hugging Face has been used for tokenization and model training. In tokenization, the sentence is converted into vector form for a model to process. To fine-tune, the 'DistilBertForSequenceClassification' pre-trained model is used for sequence classification. As there are more than one possible output classes, the 'num_labels' parameter needs to specify explicitly. The hyper-parameters required were specified using the 'TrainingArguments' library imported from transformers. The batch size is 8 entries per batch, and 5 epochs were used in training.

Evaluation Metrics:-

The sklearn fails to calculate precision, recall, and F1 score because there is more than one output for each input. The accuracy score was 1.

Deployed Model: https://huggingface.co/spaces/nikhil567/Emotion-Tracker