Course: **Natural Language Processing [NULL]** 05-September-2022 **(Spring 2022)**

Resource Person: **Muhammad Shakeel** **Course Project (Applying Transformer Models)**

Total Points: 25 (Code Descriptions) + 25 (Code) = 50

Submission Due: Wednesday, September 21, 2022

Instructions: Please Read Carefully!

- **This is an individual assignment. Everyone is expected to complete the given assignment on their own, without seeking any help from any website or any other individual.** There will be strict penalties for any work found copied from any source and the university policy on plagiarism will be strictly enforced.
- You are expected to submit this assignment as:
    a. Create a **single Python Jupyter Notebook file** for the assignment solution. The name of the file should be your student ID.
    b. Create a **PDF** file of your Notebook file. Make sure that the output of **ALL** Notebook cells is clearly visible and there are no error messages. The name of the file should be your student ID.
    c. Create a **zip file** having name as your ID. Add both files in the zip file and TURN IN against this assignment. **Do not create a .rar file.**
- Assignment is to be submitted on the **Google Classroom only**.

PROJECT DESCRIPTION [50]

NOTE: This class project is worth 10% of the total course grade.

This class project requires you to explore and study various transformer models, especially as implemented by the **Hugging Face Community** (https://huggingface.co/models). These models are also available in the **spaCy library** (https://spacy.io/api/transformer). This project requires you to apply any one of the transformer models to the given dataset for the sentiment classification task. The choice of the transform models to use is up to your own choice. You may look at using the **DistilBERT** transformer since it is a light, small, and fast version of its bigger BERT version. However, this selection is totally up to your own choice.

For this task, you have been given a large movie reviews dataset containing 25000 positive and same number of negative reviews as **train** and **test** sets in separate folders. Please study the structure of the given dataset first and understand how the data has been presented in it.

For this project, you will need to create a **Jupyter Notebook file** preferably on the Google Colab / Kaggle / or the Gradient platforms (for accessing GPU resources) that uses the **spaCy library** and performs the following operations on this corpus:

1. Please use the spaCy provided text preprocessing, such as tokenization, and data cleanup functions.
2. Make sure to use the GPU-based resources while training and testing phase of your work.
3. Select a sample of **5000** training examples and train it using the transformer of your own choosing. You should describe which transformer model you have selected and the reason for its selection. Also, you should describe how this transformer is used in code.
4. Test the training results on a sample of **500** test examples as given in the dataset.
5. Please report the type of the evaluation statistics used and their values at the test phase.
6. Vary the number of epochs during training: you can start the training from 10 epochs and increase it to 20 and report the effects on the training and test processes.
7. Make sure to thoroughly describe each code cell, starting from import statements, so that your code could be easily understood by anyone who is not familiar with the usage of transformers and clearly understands how to use them based on your descriptions.
8. Give proper headings to your descriptions.
9. Make sure that the top of your Notebook contains the course code and name, your ID and name, and the title as **DS-315: NLP Project**. Please format them appropriately.
10. Make sure to display all code cell outputs of your Notebook and convert to a PDF file so that one can easily read your complete code, their outputs, and descriptions from the PDF file. There should be no error messages in your Notebook.
11. Submit your Notebook and the resulting PDF files as one zip file. Please see the instructions on the previous page on how to give names to these files.
12. Any submission not following the given guidelines will be **heavily penalized**.

END OF PROJECT DESCRIPTION