ELL884
Assignment 3
MultiLabel Text Classification

Deadline: 15/04/2024

Task: Multi-label classification refers to a supervised learning scenario wherein a singular instance or sample may be linked to numerous labels or categorizations. For instance, named entity recognition.

Dataset: This dataset consists of an approximately 50,000 collection of research articles. Each article is described in terms of 14 labels. The dataset can be downloaded from:

https://drive.google.com/file/d/1iqk6XbNtTMVBSw3ON-uCrJ7oifEW6n4V/view?usp=sharing

The labels are mapped to actual names as follows:

```
"A": "Anatomy"
"B": "Organisms"
"C": "Diseases"
"D": "Chemicals and Drugs"
"E": "Analytical, Diagnostic and Therapeutic Techniques, and Equipment"
"F": "Psychiatry and Psychology"
"G": "Phenomena and Processes"
"H": "Disciplines and Occupations"
"I": "Anthropology, Education, Sociology, and Social Phenomena"
"J": "Technology, Industry, and Agriculture"
"L": "Information Science"
"M": "Named Groups"
"N": "Health Care"
"Z": "Geographicals"
```

Note: Only PyTorch or Tensorflow is allowed for the assignment.

Your task is to split the dataset into the ***train***, ***test*** and preferably ***validation*** datasets and train a deep learning model using the dataset to predict the class of the text. Test the model on the test dataset. You are allowed to use pre-trained language models like BERT, GPT or whatever you feel like. But keep in mind that you have to either finetune the existing model or add extra layers to it to be trained for our downstream task. *You are not allowed to use Word2Vec, GLoVE, FastText etc, to generate the embeddings and feed them into neural networks*. You are supposed to use the *title* and *abstract text* to

learn the embeddings. It would be good if you could show a comparison between the various settings you have tried; however, implementing one model completely would be sufficient as well. You are free to create additional features. Since it is a multilabel classification, you are supposed to experiment with the activation functions and loss functions.

Helper Modules:

Read File

```
import pandas as pd
data='Multi Label Text Classification Dataset.csv'
df= pd.read_csv(dataset_Name)
df.head(3)
```

	Title	abstractText	meshMajor	pmid	meshid	meshroot	A	B	C	D	E	F	G	H	I	J	L	M	N	Z
0	Expression of p53 and coexistence of HPV in pr...	Fifty-four paraffin embedded tissue sections f...	['DNA Probes, HPV', 'DNA, Viral', 'Female', 'H...	8549602	[['D13.444.600.223.555', 'D27.505.259.750.600....	['Chemicals and Drugs [D]', 'Organisms [B]', '...	0	1	1	1	1	0	0	1	0	0	0	0	0	0
1	Vitamin D status in pregnant Indian women acro...	The present cross-sectional study was conducte...	['Adult', 'Alkaline Phosphatase', 'Breast Feed...	21736816	[['M01.060.116'], ['D08.811.277.352.650.035'],...	['Named Groups [M]', 'Chemicals and Drugs [D]'...	0	1	1	1	1	1	1	0	1	1	0	1	1	1
2	[Identification of a functionally important di...	The occurrence of individual amino acids and d...	['Amino Acid Sequence', 'Analgesics, Opioid', ...	19060934	[['G02.111.570.060', 'L01.453.245.667.060'], [...	['Phenomena and Processes [G]', 'Information S...	1	1	0	1	1	0	1	0	0	0	1	0	0	0

Encode Labels as One-Hot Vectors

```
df_train['one_hot_labels'] = list(df_train[mesh_Heading_categories].values)
```

Evaluation

To avoid confusion, you will be evaluated on the performance of one of your models. The following metrics need to be computed for evaluation:

1. For each class
    a. Precision
    b. Recall
    c. F1-Score
2. Aggregate Metrics
    a. Micro Average
    b. Macro Average

Submission: Please submit the assignment by joining the following Google Classroom.