Project #5 for the Biomedical Information Retrieval Course

Name: Phan Ben

ID Student: P76127051

1. ***Dataset Processing:***
   To preprocess the dataset, we first handled the XML files and utilized the Pandas library to create dataframes for Exploratory Data Analysis (EDA). The goal was to gain insights into the data distribution, patterns, and characteristics.

2. ***Model Training:***
   - BERT Architecture: I opted for BERT as the model for our classification task and designed two main architectures:
     - BERT Model 1: **{Entity e1} [SEP] {Entity e2} [SEP] {full_sentence}**
     - BERT Model 2: **{Entity e1} [SEP] {Entity e2}**
   - Loss Function: Given the imbalance in the dataset, we employed the Focus Loss to address the issue of unbalanced data distribution.
   - Pretraining Models: I experimented with two pretraining models:
     - ***bert-base-uncased***: The default BERT model.
     - ***alvaroalon2/biobert_diseases_ner***: A model fine-tuned with biomedical data.

3. ***Evaluation:***

Model	Pretrain Model	Devel Dataset			Test Dataset		
		Recall	Precision	F1 Score	Recall	Precision	F1 Score
BERT 1	bert-base-uncased	0.8983	0.8922	0.8953	0.8724	0.9013	0.8869
BERT 1	**alvaroalon2/biobert_diseases_ner**	**0.9051**	**0.9297**	**0.9103**	**0.9048**	**0.9298**	**0.9071**
BERT 2	bert-base-uncased	0.7993	0.7594	0.7703	0.7972	0.7343	0.7557
BERT 2	alvaroalon2/biobert_diseases_ner	0.8153	0.7833	0.7919	0.7907	0.7504	0.7636

   - We assessed the performance using metrics such as Recall, Precision, and F1 Score on both the Test and Development (Devel) datasets.
   - The comparison table revealed that BERT Model 1 with the **alvaroalon2/biobert_diseases_ner** achieved the highest F1 Score on the Test dataset (**0.9071**) and Devel dataset (**0.9103**).
   - This superiority can be attributed to training on biomedical data with full sentence context.

4. ***Inference:***
   To demonstrate the model's functionality, we developed a website for running live demos. Users can input entity - entity Interaction sentences, and the system will classify them into predefined categories based on the trained BERT model.