# **Text Classification on the Web of Science Dataset using BERT**

This project uses the BERT model for text classification on the Web of Science dataset. The project includes code for downloading the dataset, preparing the data for training, creating the model, training the model, and evaluating the model’s performance.

## **Installation**

To run this project, you need to install the following Python libraries:

Python

!pip install pandas numpy matplotlib collections-extended itertools requests zipfile io torch transformers scikit-learn

Please note that to use the GPU for training the model, you need to have CUDA installed on your machine. You can check if the necessary libraries are installed and if CUDA is available by running the following code:

Python

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import collections

import itertools

import requests

import zipfile

import io

import torch

from transformers import BertTokenizer, BertForSequenceClassification

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import classification\_report

# Check if CUDA is available

if torch.cuda.is\_available():

print("CUDA is available")

else:

print("CUDA is not available. Please install CUDA to use the GPU for training the model.")



If “CUDA is available” is printed, then you are ready to proceed with running the project. If not, you will need to install CUDA to use the GPU for training the model.

# **Text Summarization**

The notebook ”text\_summarization.ipynb” runs on Google Colab machines, mostly using GPU. All the required packages are listed in the Setup chapter of the notebook. The libraries installed in the python environment are also listed in the file “requirements.txt”.

In the folder “other” you will find:

* “quillbot.py” and “scraper.py”, used to generate the ground truth summaries;
* “rouge\_pt\_bart.pkl”, a support file that contains values saved in the notebook;
* “summaries5736.txt”, the ground truth scraped using QuillBot;
* “requirements.txt”, the file listing all the env libraries version.