Configuration Manual

MSc Research Project
Data Analytics

Sourav Prabhakar Bhor
Student ID: x19231741

School of Computing
National College of Ireland

Supervisor: Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sourav Prabhakar Bhor
Student ID:	x19231741
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Majid Latifi
Submission Due Date:	16/12/2021
Project Title:	Configuration Manual
Word Count:	350
Page Count:	3

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	2nd February 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Sourav Prabhakar Bhor
x19231741

1 Prerequisite Configuration

1.1 Python Setup

- Install python on your system. Minimum required version 3.9.x. Download it from: `https://www.python.org/downloads/`
- For Instructions on how to download and install it on system visit the link: `https://docs.python.org/3/using/index.html`

1.2 PyTorch & CUDA Setup

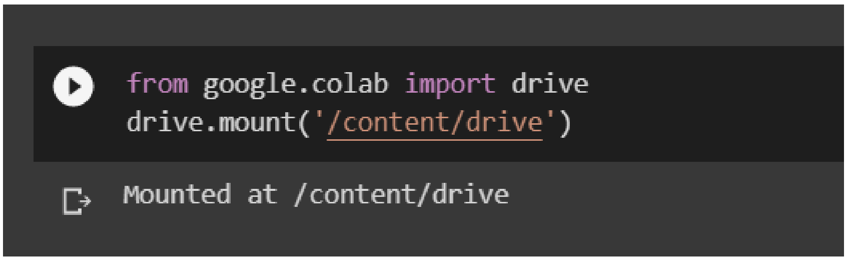
- Skip CUDA Setup step if GPU not available for your system
  - Install CUDA version from: `https://developer.nvidia.com/cuda-11-3-0-download-archive`
  - Install PyTorch version compatible with CUDA version. We use PyTorch version 1.10.1 with CUDA 11.3. Use the command:

    ```
    pip3installtorch==1.10.1+cu113torchvision==0.11.2+cu113torchaudio=
    ==0.10.1+cu113-fhttps://download.pytorch.org/whl/cu113/torch_stable.
    html
    ```
- If only CPU is available, use the below command:

  ```
  pip3installtorchtorchvisiontorchaudio
  ```

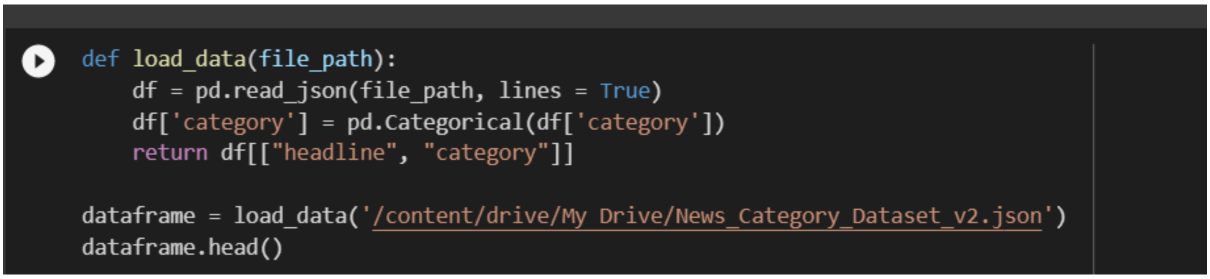
2 Running EDA and Baseline model

- Open the file with name EDA_BASIC_BERT.ipynb using Google Colab or Jupyter
- If opened using Google Colab, upload dataset News_Category_Dataset_v2.json into a google drive
- It is recommended that this code is run in Google Colab due to various computational and resource requirements of BERT.
- If opened using Jupyter:
  - Comment lines shown in Figure 1
  - Give path to the dataset News_Category_Dataset_v2.json on line shown in Figure 2



```
from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive
```

Figure 1: Google drive loading code



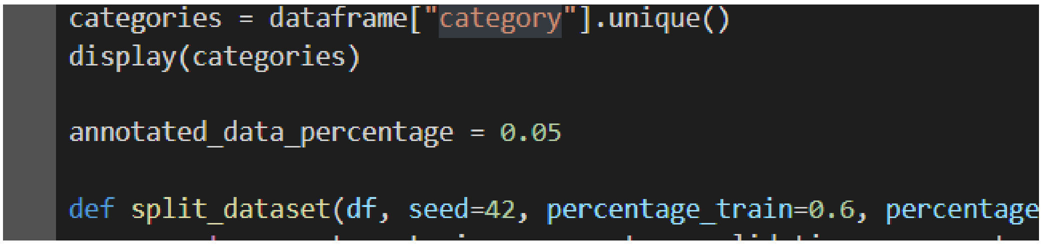
```
def load_data(file_path):
    df = pd.read_json(file_path, lines = True)
    df['category'] = pd.Categorical(df['category'])
    return df[["headline", "category"]]

dataframe = load_data('/content/drive/My Drive/News_Category_Dataset_v2.json')
dataframe.head()
```

Figure 2: Dataset loading code

- To change the annotated data percentage, change the value of the variable **annotated_data_percentage** as shown in Figure 3. It is set at 0.05 by default which is 5% as stated in the report.

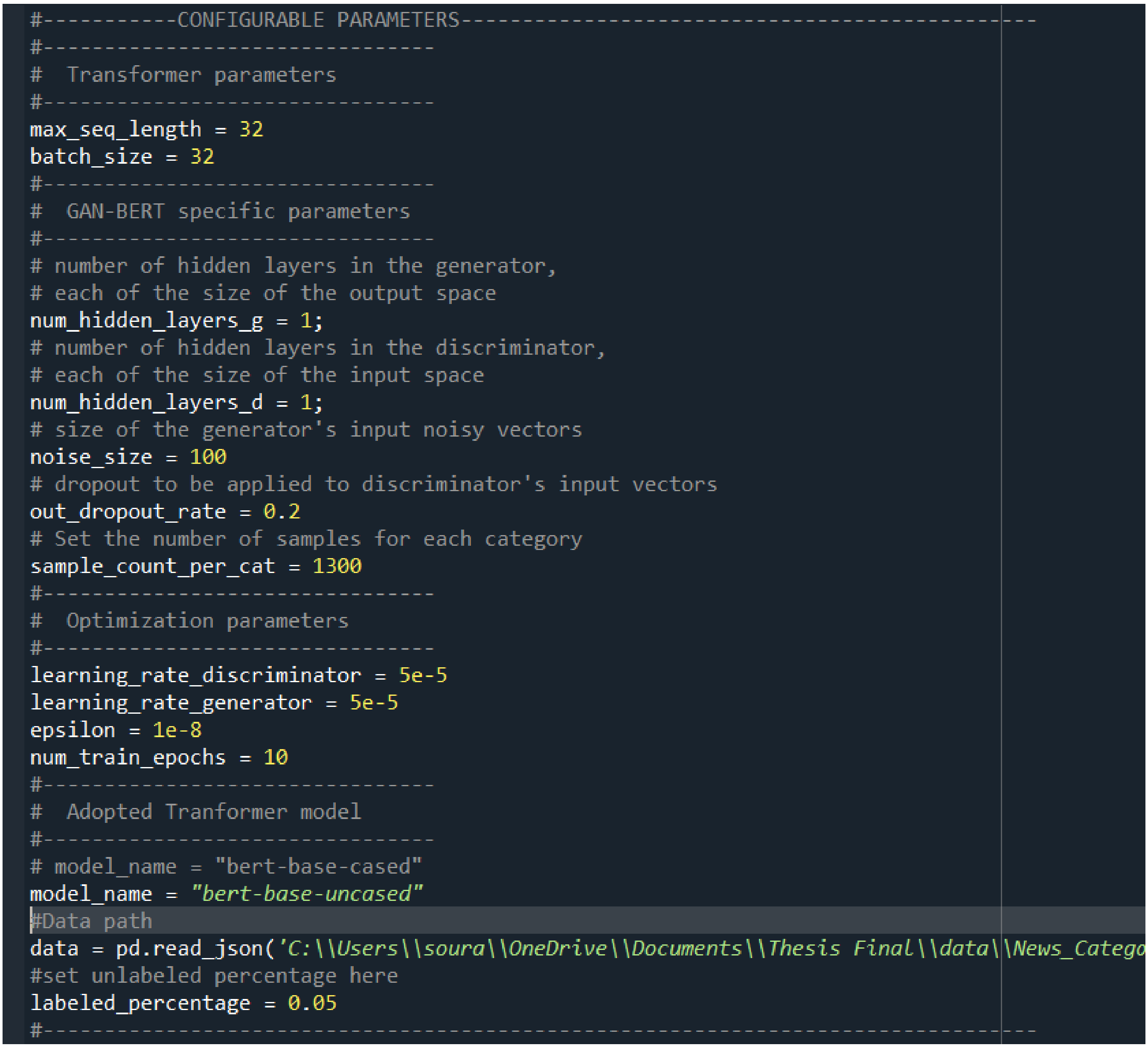
3 Running GAN-BERT

- The code is in file GAN_BERT.py
- Install Transformer package version 4.12.5 from pip.
- Set data path and all configurable parameters as shown is Figure 4. The comments describe the parameter below it.
- Run command: python GAN_BERT.py
- If re-running the code again, restart the kernel.



```
categories = dataframe["category"].unique()
display(categories)

annotated_data_percentage = 0.05

def split_dataset(df, seed=42, percentage_train=0.6, percentage
```

Figure 3: To change annotated data percentage



```
#-----------CONFIGURABLE PARAMETERS-----------------------------------------------
#--------------------------------
#  Transformer parameters
#--------------------------------
max_seq_length = 32
batch_size = 32
#--------------------------------
#  GAN-BERT specific parameters
#--------------------------------
# number of hidden layers in the generator,
# each of the size of the output space
num_hidden_layers_g = 1;
# number of hidden layers in the discriminator,
# each of the size of the input space
num_hidden_layers_d = 1;
# size of the generator's input noisy vectors
noise_size = 100
# dropout to be applied to discriminator's input vectors
out_dropout_rate = 0.2
# Set the number of samples for each category
sample_count_per_cat = 1300
#--------------------------------
#  Optimization parameters
#--------------------------------
learning_rate_discriminator = 5e-5
learning_rate_generator = 5e-5
epsilon = 1e-8
num_train_epochs = 10
#--------------------------------
#  Adopted Tranformer model
#--------------------------------
# model_name = "bert-base-cased"
model_name = "bert-base-uncased"
#Data path
data = pd.read_json('C:\\Users\\soura\\OneDrive\\Documents\\Thesis Final\\data\\News_Catego
#set unlabeled percentage here
labeled_percentage = 0.05
#---------------------------------------------------------------------------------
```

Figure 4: Configurable parameters for GAN-BERT