National College of Ireland

Configuration Manual

MSc Research Project
Master of Science in Artificial Intelligence

Sharanya Neelakanti
Student ID: X23138220

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Sharanya Neelakanti……

Student ID: ……X23138220……

Programme: ……MSc in Artificial Intelligence…… **Year:** …2023-2024…

Module: ……MSc Research Practicum……

Lecturer: ……Rejwanul Haque……

Submission Due Date: …… 16 September 2024……

Project Title: …Comparative Analysis of Transformer Models for Multi-Class Text Classification……

Word Count: ……996…… **Page Count:** ……8……

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: ……Sharanya Neelakanti……

Date: ……16 September 2024……

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

Sharanya Neelakanti
Student ID: X23138220

1 INTRODUCTION

This is an inference of sentiment analysis implemented using a battery of fine-tuned pre-trained language models: BERT, RoBERTa, XLNet, DistilBERT, and GPT-3.5 on the GoEmotions dataset. The system shall be made up of the following three main text classifications: positive, negative, and neutral sentiments. Key features of the project include:

- Loading pre-trained models for inference
- Evaluating model performance on a test dataset
- Predicting sentiment for single text inputs
- Handling both traditional transformer models and the OpenAI GPT-3.5 API
- Configs of Models and All

1.1 Test Dataset:

- Source: GoEmotions dataset (preprocessed)
- Classes: Mapped to 3 main classes (positive, neutral, negative)

2 Connecting Google colab

2.1 Open Google Colab

- Go to https://colab.research.google.com/
- Create a new notebook or open an existing one

2.2 Connect to GPU Session

- Click on "Runtime" in the top menu
- Select "Change runtime type"
- In the pop-up window, set "Hardware accelerator" to "T4 GPU"
- Click "Save”

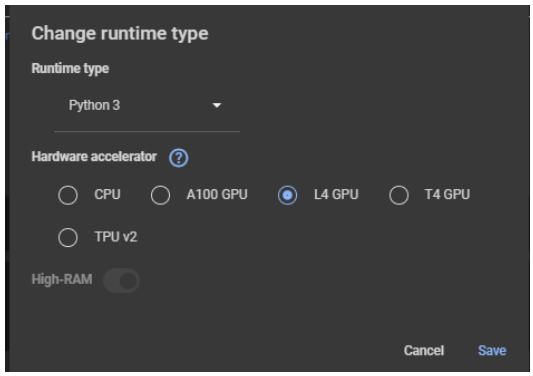

Figure 1 GPU Selection

2.3 Mount Google Drive

- Run the following code to mount your Google Drive

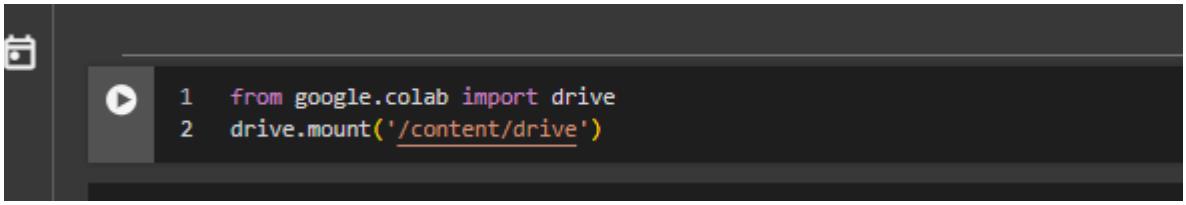

Figure 2 Google drive Mount

3 Setup and Configuration

3.1 Directory Structure

Create the following directory structure, replacing "PROJECT_NAME" with a name of your choice

```
/content/drive/MyDrive/PROJECT_NAME/
├── Dataset/
├── BERT_MODEL/
├── ROBERTA_MODEL/
├── XLNET_MODEL/
└── DISTILBERT_MODEL/
```

Figure 3 Directory Structure

3.2 Dataset Upload

Upload the preprocessed GoEmotions dataset:

- File: 'go_emotions_merged_augmented.csv' (or your specific filename)
- Upload to: '/content/drive/MyDrive/PROJECT_NAME/Dataset/'

3.3 Model Upload

Upload the pre-trained models to their respective directories:

- BERT: Upload to '/content/drive/MyDrive/PROJECT_NAME/BERT_MODEL/'
- RoBERTa: Upload to '/content/drive/MyDrive/PROJECT_NAME/ROBERTA_MODEL/'
- XLNet: Upload to '/content/drive/MyDrive/PROJECT_NAME/XLNET_MODEL/'
- DistilBERT: Upload to '/content/drive/MyDrive/PROJECT_NAME/DISTILBERT_MODEL/'

3.4 Model Paths

3.4.1 Update the MODEL_PATHS dictionary in your code to reflect your chosen directory structure:

```
MODEL_PATHS = {
    'BERT': '/content/drive/MyDrive/PROJECT_NAME/BERT_MODEL/best_model_BERT',
    'RoBERTa': '/content/drive/MyDrive/PROJECT_NAME/ROBERTA_MODEL/best_model_RoBERTa',
    'XLNet': '/content/drive/MyDrive/PROJECT_NAME/XLNET_MODEL/best_model_XLNet',
    'DistilBERT': '/content/drive/MyDrive/PROJECT_NAME/DISTILBERT_MODEL/best_model_DistilBERT',
    'GPT-3.5': "ft:gpt-3.5-turbo-0125:personal:nci-thesis:9tKr25Ma"
}
```

Figure 4 Model Paths

3.4.2 Update the dataset path as well like below

```
    print("\n")

def main():
    # Load your dataset
    data = pd.read_csv('/content/drive/MyDrive/PROJECT_NAME/Dataset/go_emotions_merged_augmented.csv')
    # Print the first few rows and column names for debugging
    print("Dataset columns:", data.columns)
    print("\nFirst few rows of the dataset:")
    print(data.head())

    while True:
        print("\nChoose an option:")
        print("1. Evaluate all models on the dataset")
        print("2. Evaluate all models on a limited number of rows")
        print("3. Get predictions for a single input")
        print("4. Exit")

        choice = input("Enter your choice (1-4): ")

        if choice == '1':
            evaluate_models(data)
```

Figure 5 Dataset Path change

3.4.3 OpenAI API Configuration

```
# Securely set your OpenAI API key
os.environ['OPENAI_API_KEY'] = 'sk-None-QfeHnc5xeLHlHfLT2pClT3BlbkFJQBZJwd9uH7h5TEOD19ji'
client = openai.OpenAI()
```

Figure 6 OpenAI API Key

Note: The OpenAI API key provided in the code is specific to the fine-tuned model for this task. Do not change this key.

3.4.4 Model Loading

- BERT, RoBERTa, XLNet, and DistilBERT: Loaded using their respective tokenizers and model classes from the transformer's library.
- GPT-3.5: Accessed via OpenAI API using the provided fine-tuned model ID.

3.5 Library Installation

- Before running the inference, ensure all required libraries are installed. You can install them using pip

Figure 7 Library installation

4 Inference

The system provides three main functionalities for inference:

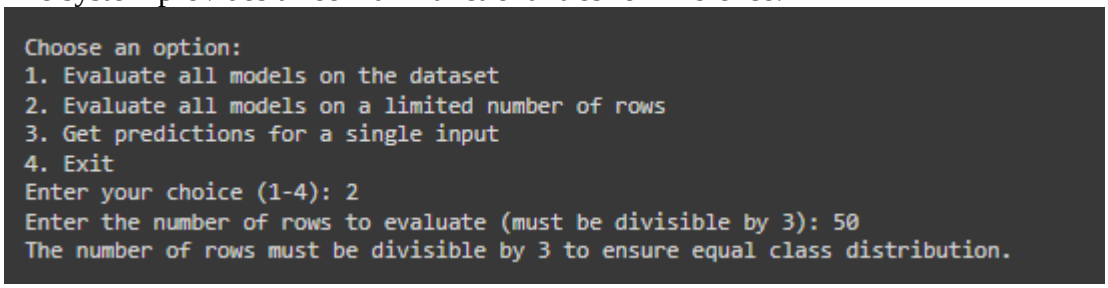

Figure 8 Options

4.1 Evaluate all models on the entire test dataset:

- Use option 1 in the main menu
- Processes the entire dataset and provides classification reports and confusion matrices for each model

4.2 Evaluate all models on a limited number of samples:

- Use option 2 in the main menu
- Enter the number of samples to evaluate (must be divisible by 3 for balanced class distribution)
- Processes a subset of the data and provides classification reports and confusion matrices for each model

4.3 Get predictions for a single input:

- Use option 3 in the main menu
- Enter the text to classify
- Preprocesses the text and provides predictions from all models

```
Choose an option:
1. Evaluate all models on the dataset
2. Evaluate all models on a limited number of rows
3. Get predictions for a single input
4. Exit
Enter your choice (1-4): 3
Enter the text to classify: NCI is a Good Girl;
Input text: NCI is a Good Girl;
Preprocessed text: nci is a good girl
BERT prediction: positive
RoBERTa prediction: positive
XLNet prediction: positive
DistilBERT prediction: positive
GPT-3.5 prediction: positive
```

4.4 To run the inference:

1. Ensure all required libraries are installed
2. Create the directory structure as described in section 3.1
3. Upload the dataset and pre-trained models as described in sections 3.2 and 3.3
4. Set up the OpenAI API key as described in section 3.5
5. Run the main() function
6. Choose the desired option from the menu and follow the prompts

Note: This code is designed for inference only. It assumes that the models have been previously fine-tuned on the GoEmotions dataset and saved. Ensure that the model paths are correct, and the pre-trained models are available before running the inference.

References

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1910.01108

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1810.04805

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1906.08237

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1907.11692

Latif, E., & Zhai, X. (2023). Fine-tuning ChatGPT for Automatic Scoring. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.10072

https://huggingface.co/datasets/mrm8488/goemotions