Detection AI-Generated Text

1. Abstract:

Recent progress in pre-trained neural language models has significantly improved the performance of many natural language processing (NLP) tasks. Since 2018, we have seen the emergence of a wide range of transformer-based pre-trained language models (PLMs), such as GPT (Radford et al., 2019; Brown et al., 2020), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019c), XLNet (Yang et al., 2019), UniLM (Dong et al., 2019), ELECTRA (Clark et al., 2020), T5 (Raffel et al., 2019), ALUM (Liu et al., 2020), ALUM (Liu et al., 2020), StructBERT (Wang et al., 2019c), and ERINE (Sun et al., 2019). These PLMs are fine-tuned with task-specific labels and create a new state of progress in many natural language processing (NLP) tasks such as generating texts in a style like what humans write, therefore, it became difficult to know the author of a text.

Part of this project is a solution to this problem—an attempt to preserve the rights of those who wrote the text and reduce the misuse of previously trained models. By building a model capable of distinguishing between texts and their writing sources using transformer-based pre-trained language models.

2.Data and Exploratory data analysis (EDA):

Data was collected from various sites to reach about **1.4** million and was as follows: The texts for humans were approximately **1.03** million and the texts for LLMs were approximately **377** thousand pieces of data, but after some preprocessing and analysis of the data and making the data equal and appropriate for this task, an attempt of **31** thousand was made, and they were divided into **28** thousand (train and validation) and **2,730** rows to test the data.

- LLM Generated Essays for the Detect AI Comp! :
    - This dataset contains 700 LLM generated essays in total. generated 500 of these essays with gpt-3.5-turbo and 200 with gpt-4.
- daigt data - llama 70b and falcon180b:
    - Contain 4 files :
        - Llama falcon v3 : contains 7000 LLM generated essays.
        - Llama 70b v2 : contains

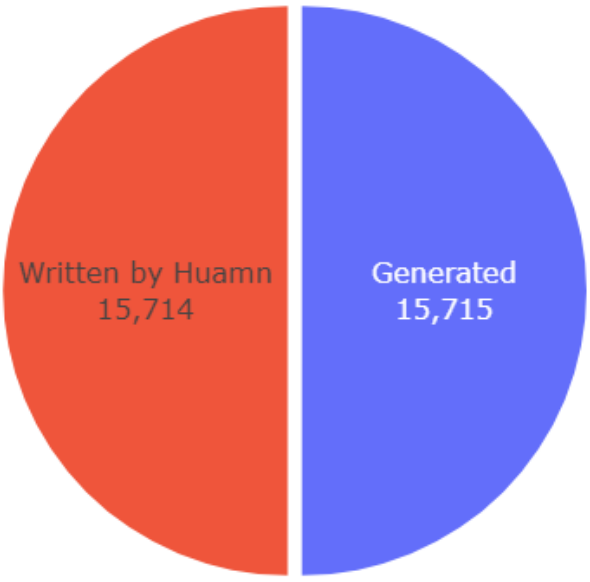

Figure 1 Data distribution

1172 LLM generated essays.
    - Llama 70b v1 : contains 1172 LLM generated essays.
    - Falcon 180b v1 : contains 1055 LLM generated essays.
- persuade corpus 2.0 :
    - the PERSUADE 2.0 corpus comprises over 25,000 argumentative essays produced by 6th-12th grade students in the United States for 15 prompts on two writing tasks. (https://github.com/scrosseye/persuade_corpus_2.0)

- DAIGT | External Dataset:
    - this dataset provides 2421 student generated texts and 2421 AI generated texts .

Figure 2: Source of data generated in Data

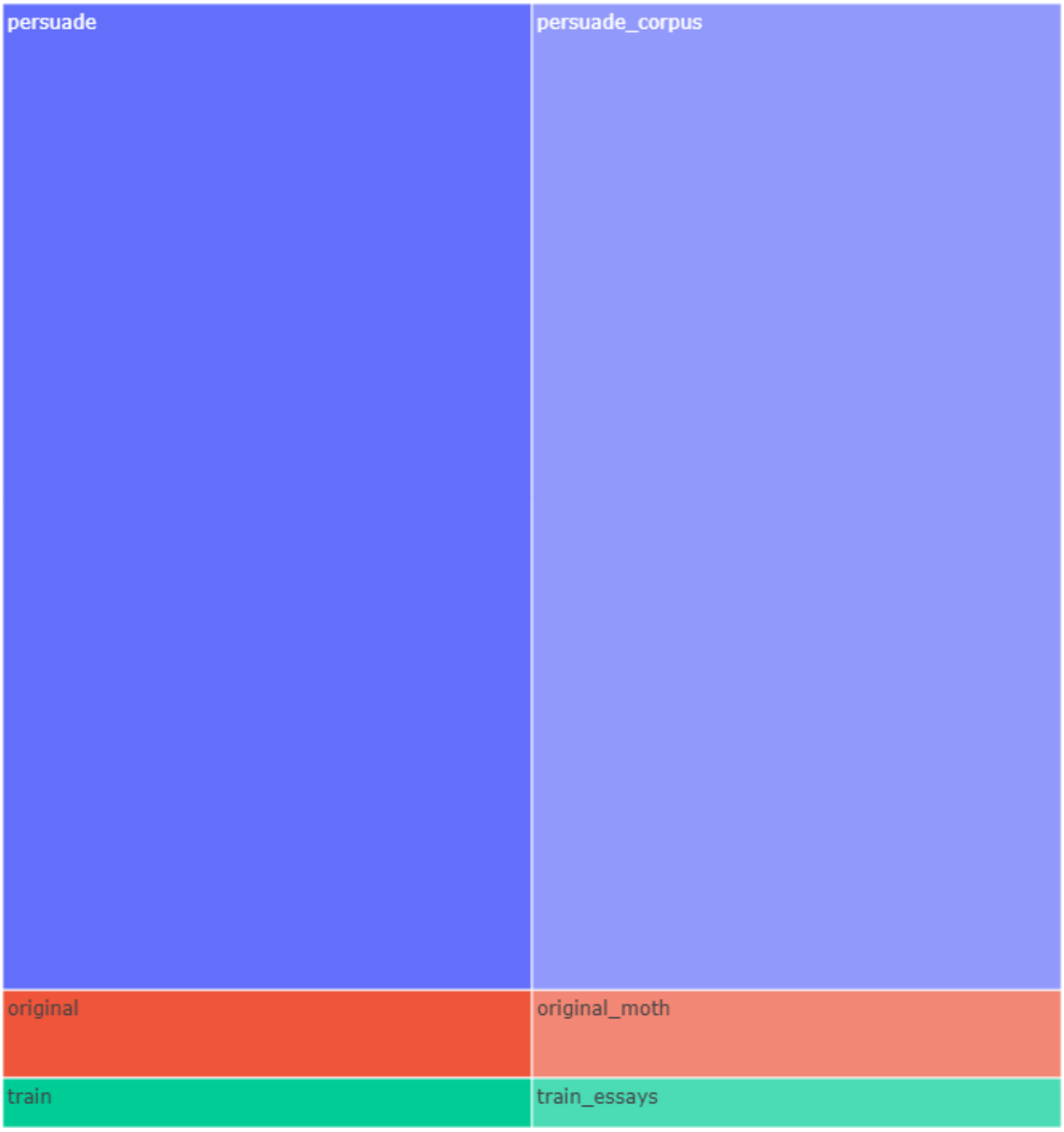

Figure 3: Source of data human in Data

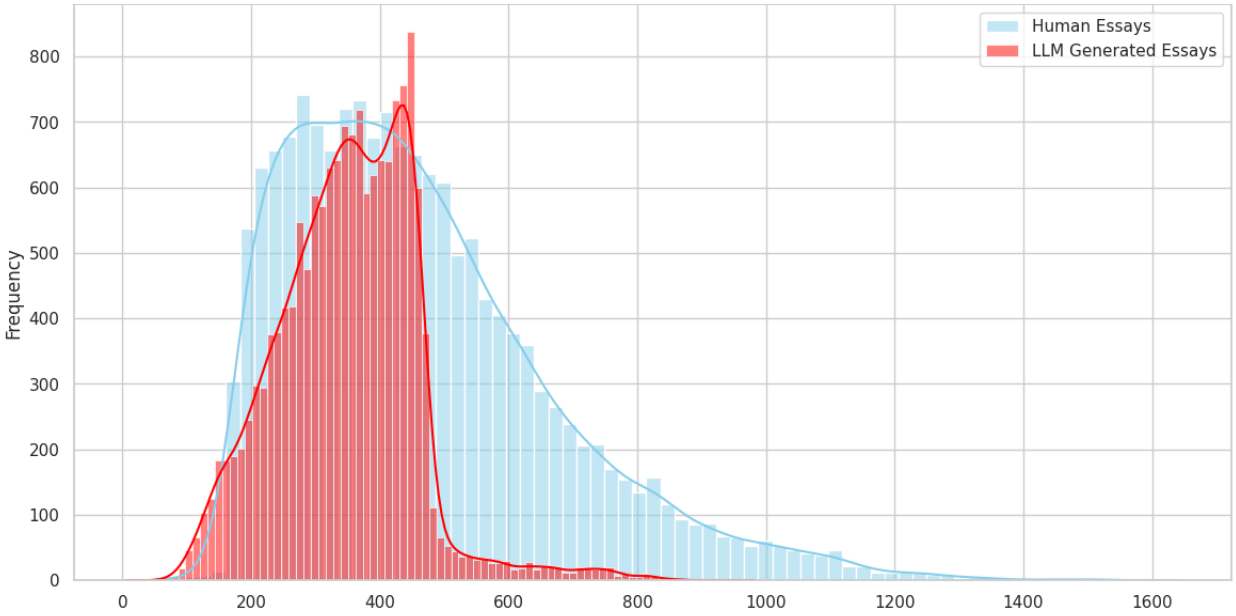

Figure 4:Histogram of data (tokens length)

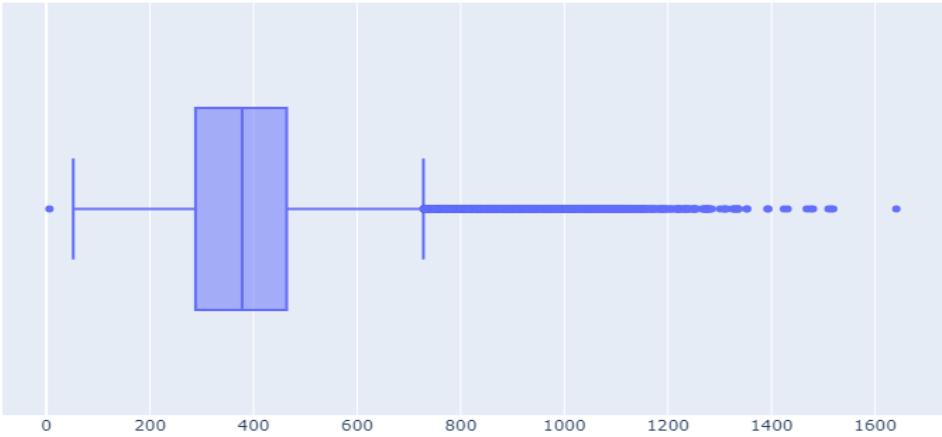

Figure 6:Boxplot on data

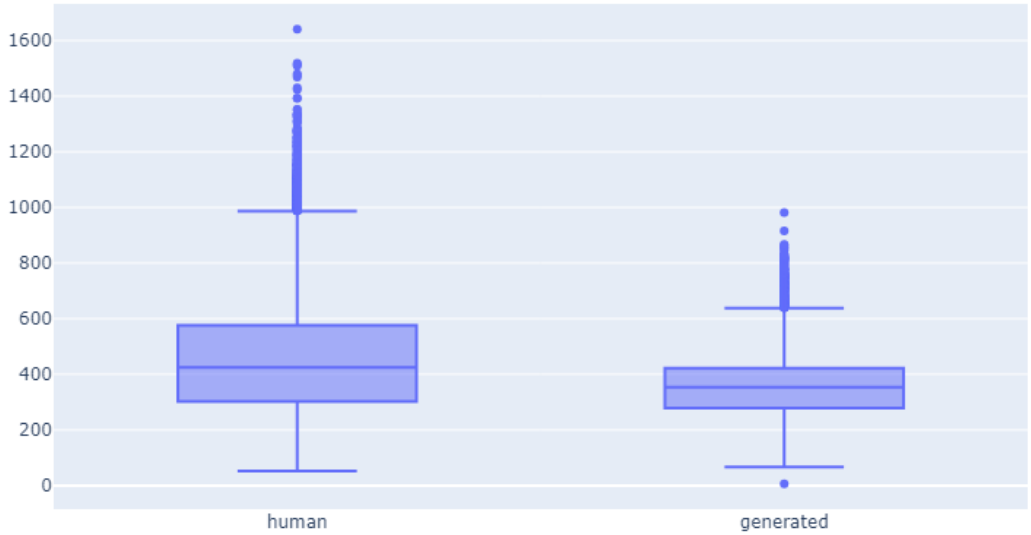

Figure 5: Boxplot based on Class

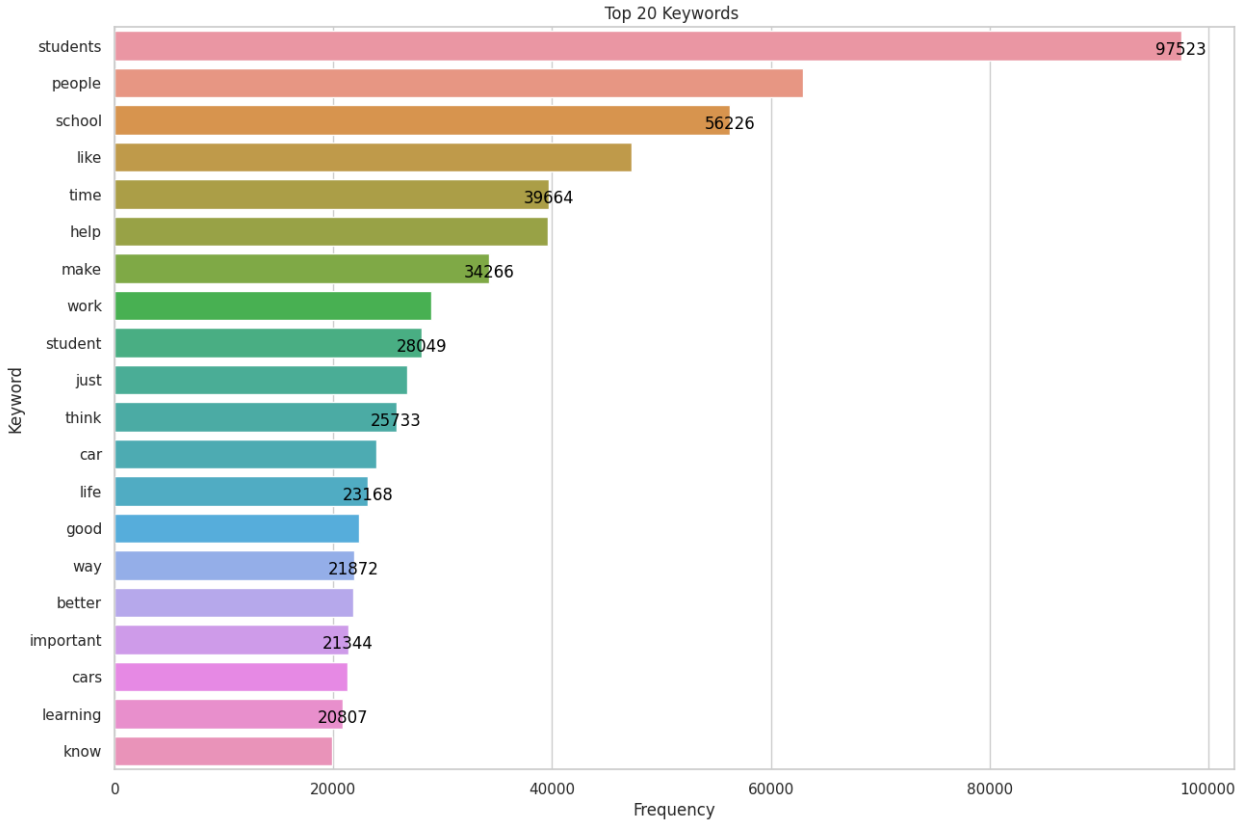

Figure 7:Common words in data

3.Models:

4 models were trained to use the best ones, as follows:

3.1BERT:

BERT [i] is a pre-trained language model developed by Google. It's widely used for various NLP tasks such as text classification. BERT Architecture utilizes a multi-layer bidirectional transformer encoder. It captures contextual relationships between words in each text by processing the entire input sequence at once. BERT Tokenizers uses WordPiece tokenization, breaking down words into subwords and representing them with corresponding embeddings.

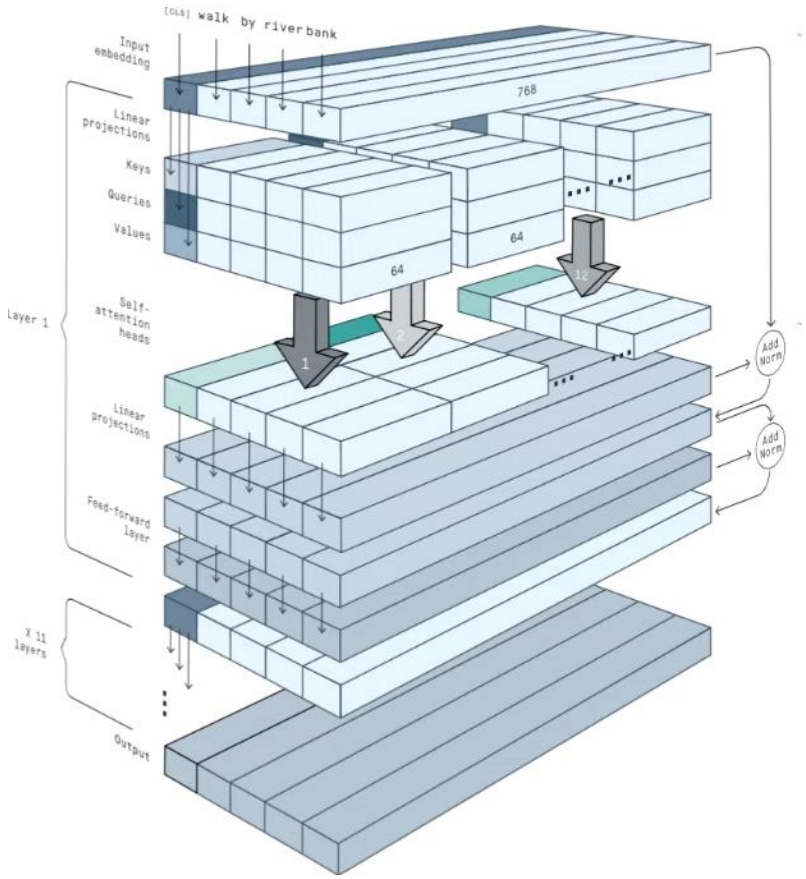

Figure 8: BERT Encoder Architecture

The data was prepared using the BERT tokenizer on Transformer library , and the max length for token is 512 tokens, and finetune model .

Training Process:

Model training with **2** epochs, set evaluation step with **100** steps , **128** batch size, use *ADAMw_torch* optimizer with learning rate equal **3e-5** with *cosine* learning rate scheduler and use *wandb* to report and Training process.

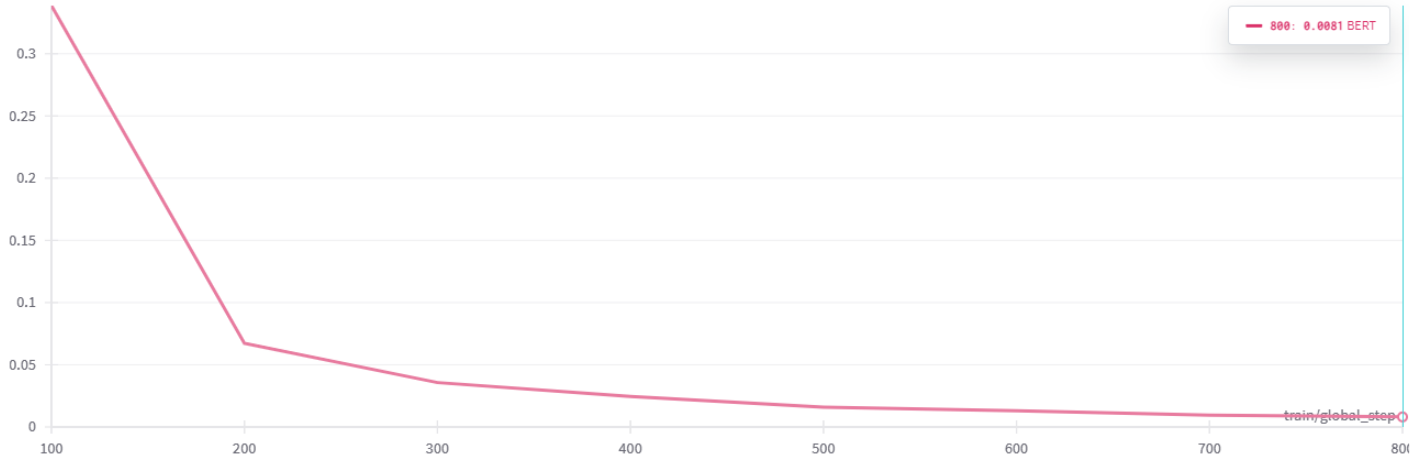

Figure 9 : Train/Loss BERT

Then , save model in HuggingFace can you use it through this link [BERT-Finetune-DAIGT.]

Evaluation Process:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	1365
1	0.99	1.00	0.99	1365
accuracy			0.99	2730
macro avg	0.99	0.99	0.99	2730
weighted avg	0.99	0.99	0.99	2730

Figure 10: Classification report BERT

- 0 refers to Written by human.
- 1 refers to Generated text.

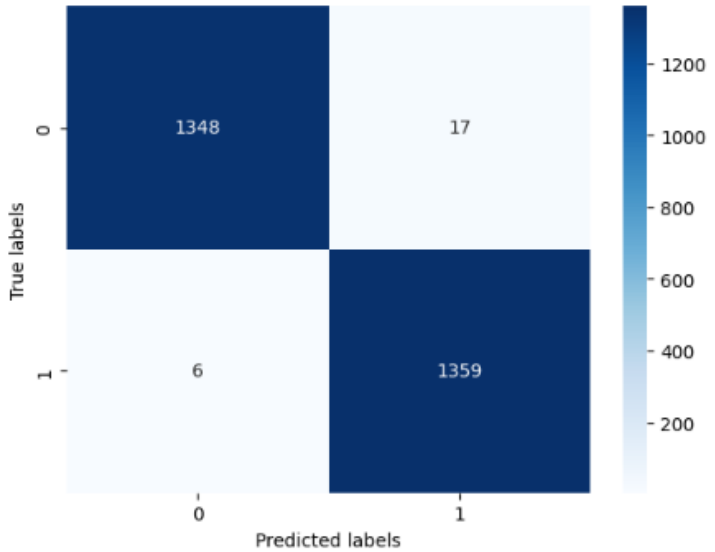

Figure 11 :Confusion matrix BERT

3.2 DistilBERT:

DistilBERT[ii] is a distilled version of BERT developed by Hugging Face. It aims to retain most of BERT's performance while being smaller and faster. DistilBERT Architecture retains the same Transformer-based architecture as BERT but with fewer layers and reduced parameters. It is distilled from the BERT model through knowledge distillation. DistilBERT Tokenizers uses the same WordPiece tokenization as BERT.

Training Process: Same train parameters BERT training process.

Figure 12: Train/Loss DistilBERT

Then , save model in HuggingFace can you use it through this link DistilBERT-Finetune-DAIGT.

Evaluation Process:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1365
1	1.00	0.99	0.99	1365
accuracy			0.99	2730
macro avg	0.99	0.99	0.99	2730
weighted avg	0.99	0.99	0.99	2730

Figure 13:Classification report DistilBERT

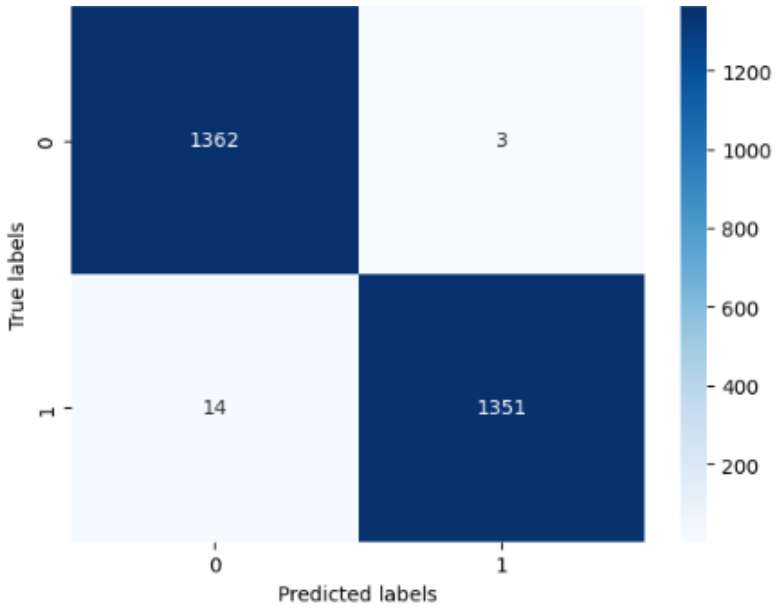

Figure 14:Confusion matrix DistilBERT

3.3 RoBERTa:

RoBERTa (Robustly optimized BERT approach)[iii] is an optimized version of BERT developed by Facebook AI. It improves BERT's performance by training on more data and with longer sequences. RoBERTa's architecture is like BERT, employing a Transformer encoder. However, it incorporates modifications such as dynamic masking and removing the next sentence prediction task during training. RoBERTa Tokenizers uses the same WordPiece tokenization as BERT.

Training Process: Same train parameters BERT training process.

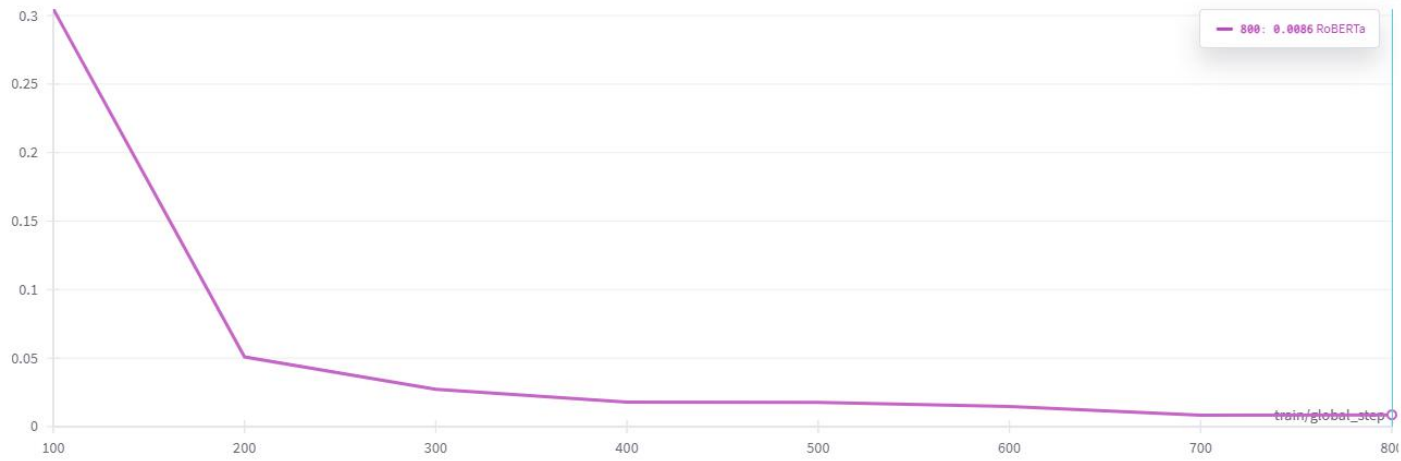

Figure 15:Train/Loss RoBERTa

Then , save model in HuggingFace can you use it through this link RoBERTa-Finetune-DAIGT.

Evaluation Process:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	1365
1	0.99	1.00	0.99	1365
accuracy			0.99	2730
macro avg	0.99	0.99	0.99	2730
weighted avg	0.99	0.99	0.99	2730

Figure 16: Classification Report RoBERTa

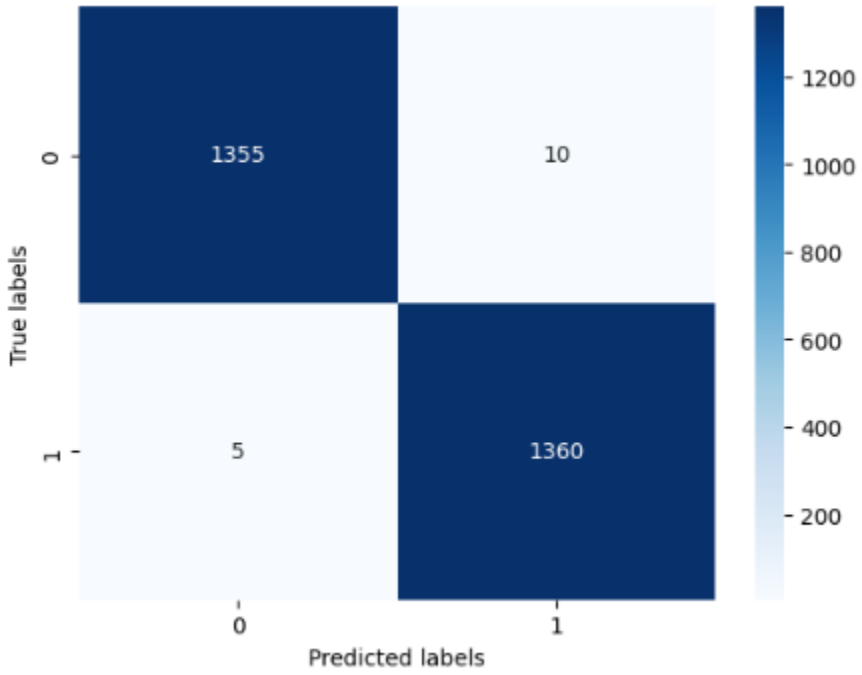

Figure 17:Confusion matrix RoBERTa

3.4 DeBERTa :

DeBERTa (Decoding-enhanced BERT with disentangled attention)[iv] is an extension of BERT designed to improve decoding efficiency and performance in downstream tasks. DeBERTa Architecture introduces disentangled attention mechanisms to better capture relationships between tokens. It also utilizes adaptive SoftMax and relative position representations for improved performance. DeBERTa Tokenizers employs the same WordPiece tokenization as BERT and RoBERTa.

Training Process: Same train parameters BERT training process.

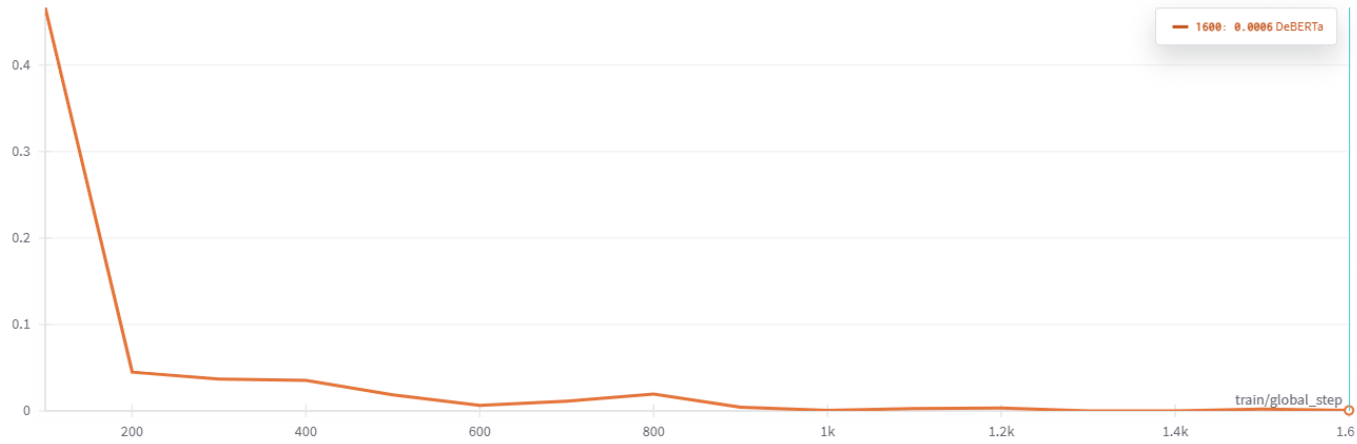

Figure 18: Train/Loss DeBERTa

Then , save model in HuggingFace can you use it through this link DeBERTa-Finetune-DAIGT.

Evaluation Process:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1365
1	1.00	1.00	1.00	1365
accuracy			1.00	2730
macro avg	1.00	1.00	1.00	2730
weighted avg	1.00	1.00	1.00	2730

Figure 19:Classification Report DeBERTa

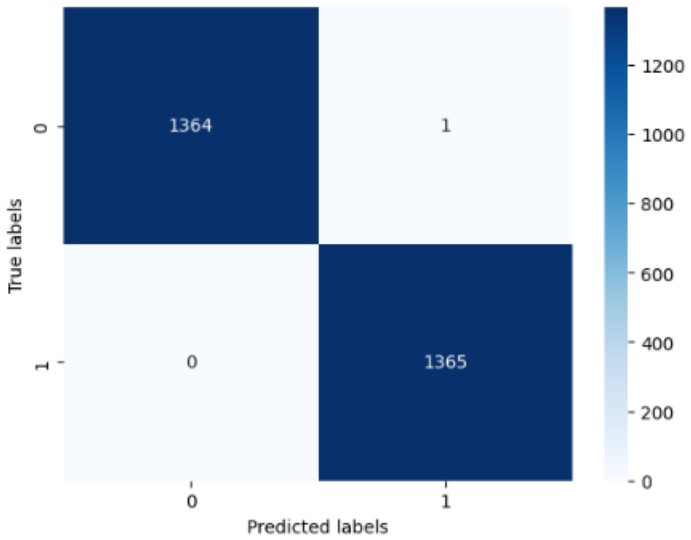

Figure 20: Confusion matrix DeBERTa

4. Methodology :

Some data is collected by us and Kaggle for use in human-evaluation. The results were close, Given that both DeBERTa and RoBERTa demonstrated promising results, particularly RoBERTa in identifying human-written text, we explored ensemble learning techniques to leverage the strengths of both models and potentially enhance verification accuracy. We implemented two ensemble methods:

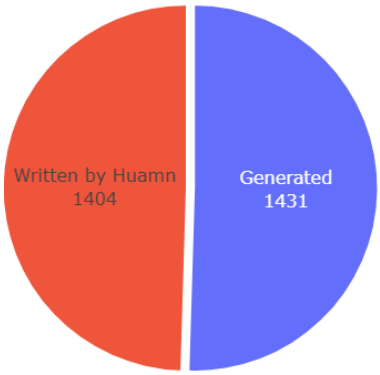

Figure 21: Data Evaluation distribution

1. **Averaging Probabilities**: This method focused on the individual model outputs – the probabilities assigned to each text being human-written or LLM-generated. We calculated the ensemble prediction by averaging the corresponding probabilities from DeBERTa and RoBERTa for each category.

2. **Feed-forward Model:** We constructed a new feed-forward neural network specifically for this task. This model took the individual model probabilities from DeBERTa and RoBERTa (human-written and LLM-generated) as input features alongside the actual ground truth labels (human-written or LLM-generated) as the target variable. The feed-forward model was then trained to learn a mapping between the combined probabilities and the correct classifications.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
BERT	*0.994356*	*0.990978*	*0.997904*	*0.994429*	*0.994322*
RoBERTa	*0.995767*	*0.998406*	*0.997904*	*0.995816*	*0.995747*
DistilBERT	*0.992240*	*0.995081*	*0.989518*	*0.992292*	*0.992266*
DeBERTa	*0.998942*	*0.997908*	*1.000000*	*0.998953*	*0.998932*
Averaging Probabilities (RoBERTa - DeBERTa)	*0.983069*	*0.998558*	*0.967855*	*0.982967*	*0.983215*
Feed-forward Model(RoBERTa -DeBERTa)	***0.999647***	***0.999302***	***1.000000***	***0.999651***	***0.999644***

Table 1: Benchmark

4.1 Final architecture model :

Based on *Table 1* in above , The feed-forward model yielded the most accurate results for text verification.

Architecture:

The model relies on a feed-forward neural network with 3 layers:

- **Input Layer**: This layer receives combined features representing the aggregated probabilities from both DeBERTa and RoBERTa models .
- **Hidden Layer:**
  This layer creates a linear transformation from the input size (which depends on how the probabilities are combined) to 32 hidden units. Applies the ReLU (Rectified Linear Unit) activation function to introduce non-linearity and improve model learning capacity.
- **Output Layer**:
  Creates a linear transformation from 32 hidden units to a single output unit, Applies the sigmoid activation function to map the output value between 0 and 1, representing the probability of the text being human-written.

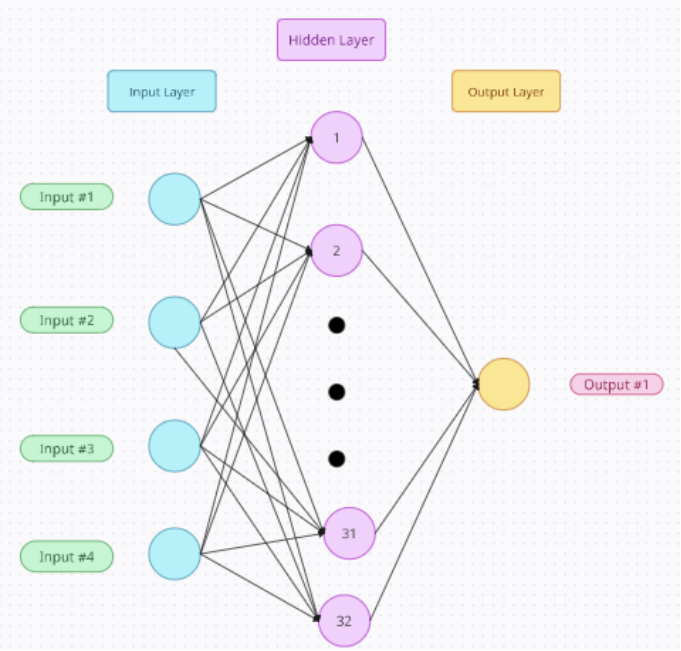

Figure 22: FF-architecture

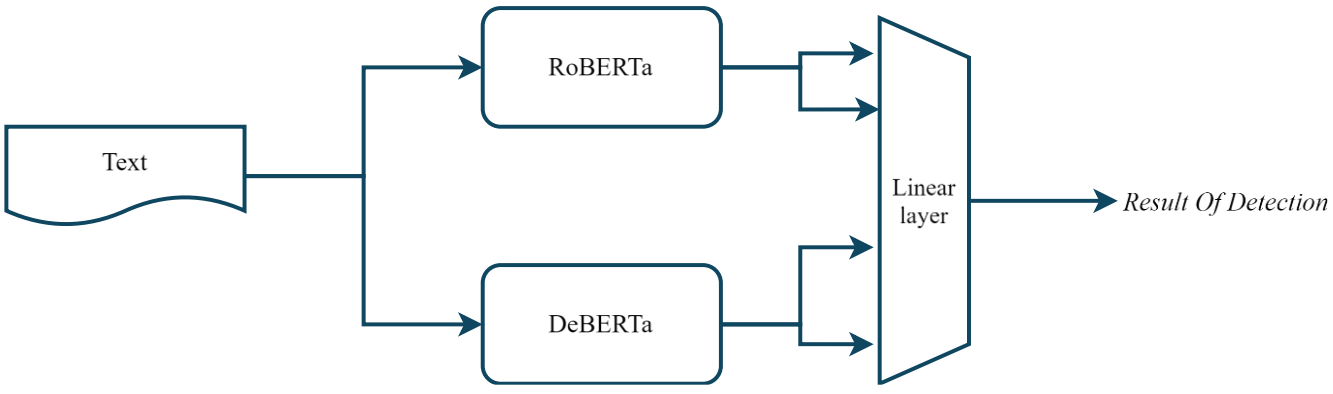

Figure 23:Final architecture model

4.2Example :

This text from Gemini about LLMs (29 tokens):

Large language models, or LLMs, are computer programs that gobble up massive amounts of text and code. They use this knowledge to mimic human language, allowing them to write different kinds of creative content, translate languages, and even answer your questions in a way that seems almost human-like. It's pretty cool stuff!

- FeedForward-RoBERTa-DeBERTa Model Prediction: *Generated*

Average Attention Scores Across All Layers:

- **RoBERTa:**

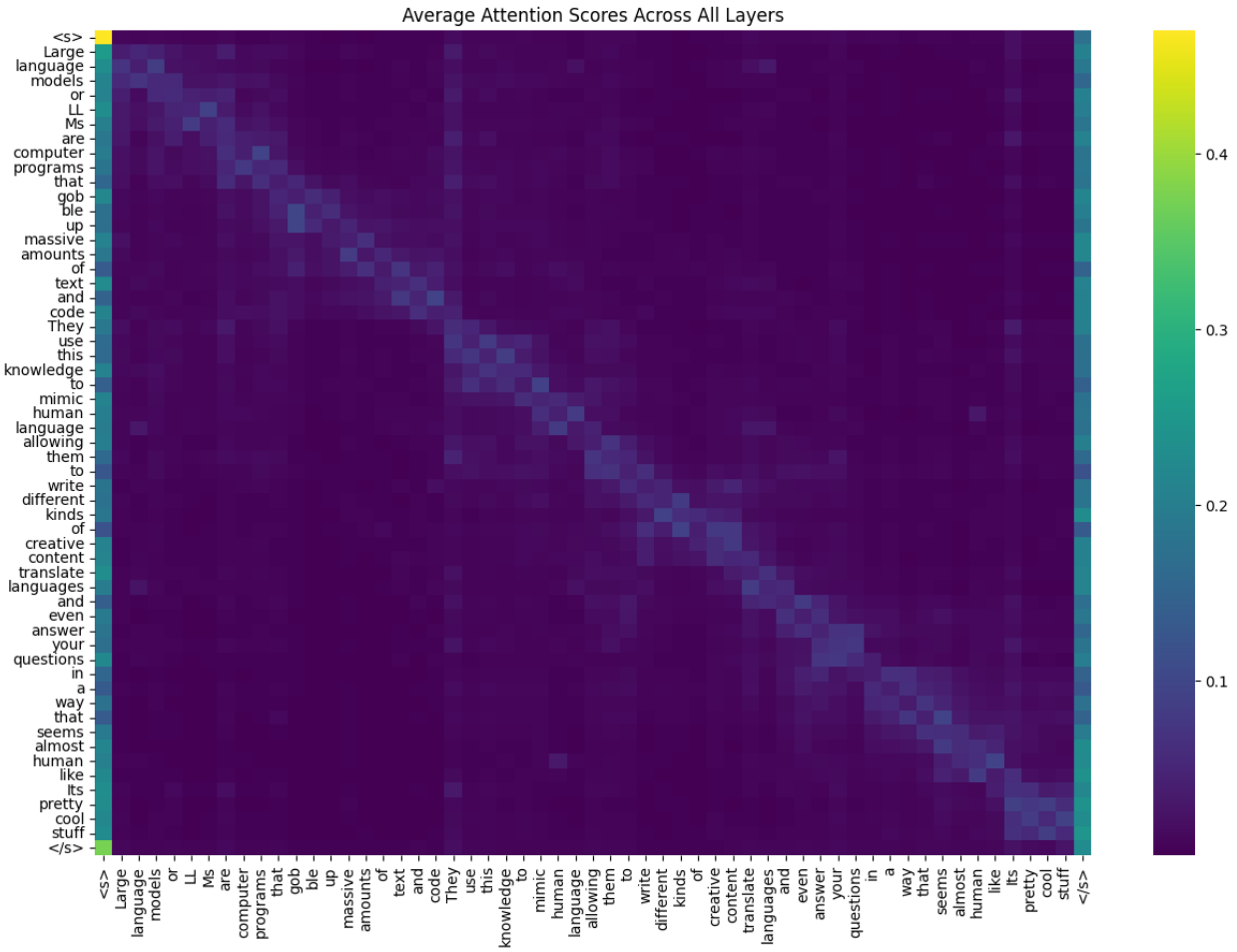

Figure 24: Attention RoBERTa

- **DeBERTa:**

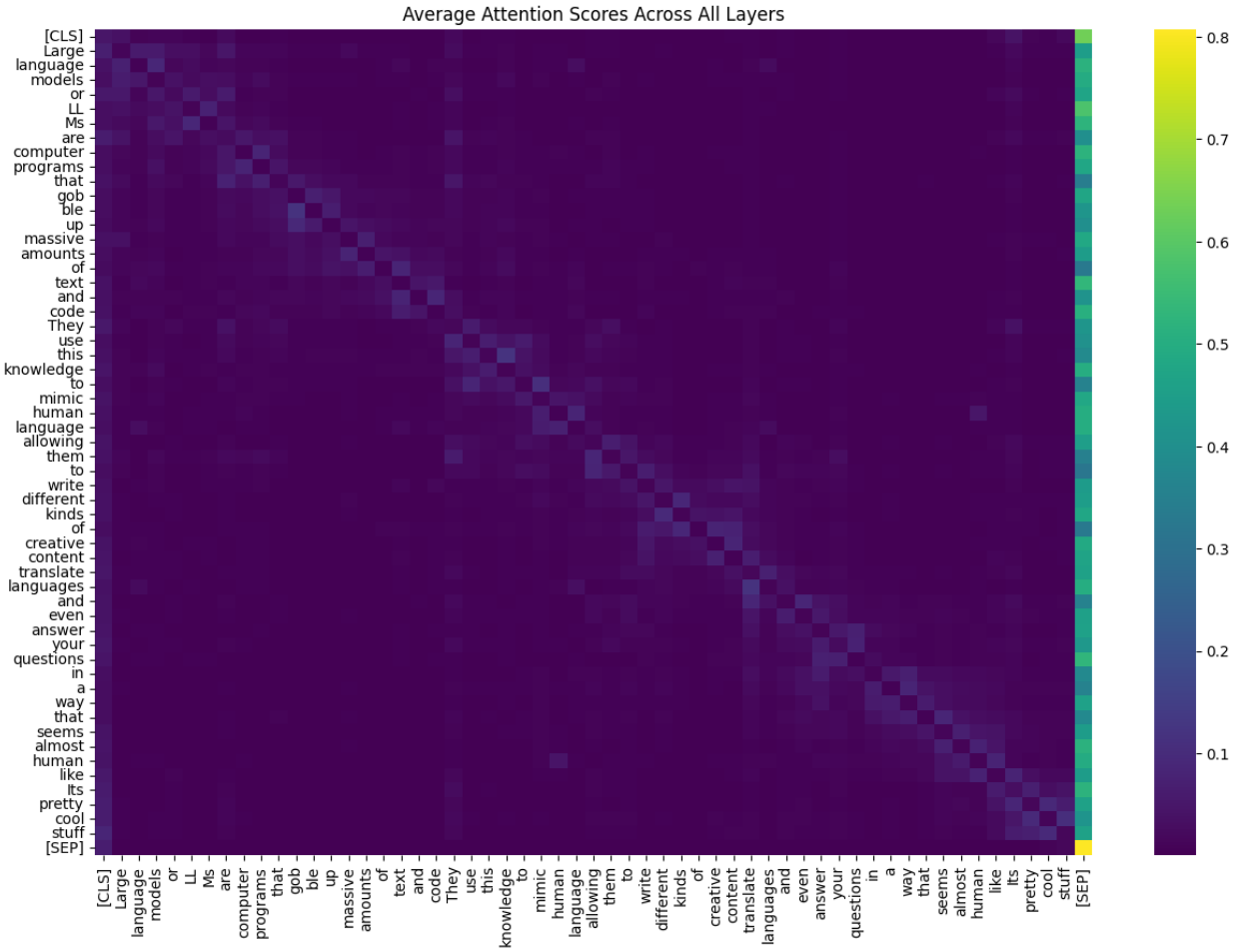

Figure 25:Attention DeBERTa

5.Reference:

[i] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[ii] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[iii] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[iv] He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.