Automatic Detection of Machine-Generated Text

Mohit Aggarwal, Aviral Agrawal, Harsh Goyal,
Georgia Institute of Technology,
Atlanta, USA
{mohit7, aagrawal433, hgoyal34, } @gatech.edu

Abstract

We will be undertaking a significant endeavour in the field of Natural Language Processing (NLP) and Machine Learning (ML) by exploring a novel SemEval-2024 conference task. The project revolves around developing a system capable of automatically detecting machine-generated text within a given document. This initiative is crucial due to the increasing prevalence of large language models (LLMs) like ChatGPT, Davinci, BLOOMz, and Dolly, which have brought about both remarkable advances and potential challenges in the generation and dissemination of text-based content. This project aims to contribute to the ongoing discourse surrounding the responsible use of large language models and their potential impact on various domains. Through the development of a robust automatic detection system, we seek to address concerns related to misinformation, educational integrity, and ethical considerations associated with the proliferation of machine-generated text on various platforms across different domains. During our detailed experimentation and results analysis, we observe that DeBERTa-v3-large outperforms the baseline model for Task A monolingual setting, whereas the DistilBERT and DeBERTa-v3-large perform comparably to the baseline models for Task A multilingual setting, Task B and Task C. All the research work done for the project has been done from scratch including the baseline models referred from the reference research paper for datasets.

1. Introduction

The project aims to achieve three primary objectives. Subtask A involves creating a model for binary classification, distinguishing between human-written and machine-generated text in both monolingual and multilingual sources. Subtask B aims to build a system to identify the source or author of a text, differentiating between texts generated by specific language models and those written by humans. Subtask C focuses on developing a mechanism to detect transition boundaries within mixed texts, indicating shifts between human-written and machine-generated content.

Detecting machine-generated text is pivotal for trust, legality, and safeguarding users. It preserves authenticity, shields against misinformation, and ensures compliance with laws. Crucially, it protects consumers from fraud, upholds intellectual property rights, and fosters fair competition. By curbing manipulation in public discourse, it promotes ethical AI use and innovation. Preserving accuracy in information and content creation, it bolsters user confidence and decision-making. These detection systems are vital safeguards against deceptive practices, maintaining a trustworthy and ethical online space.

Since the dataset has already been provided to us by the task organizers, we need not put a lot of effort into the collection of the data and rather our main objective would be to explore a novel approach to tackle this research problem. Please refer to the dataset section for a detailed description.

2. Related Work

The paper by Mitrovic et Al investigates if machine learning models can differentiate between original human-generated text and text generated by ChatGPT, particularly focusing on short texts [9]. To achieve this, the researchers study online reviews, using two experiments: one with custom ChatGPT queries and another where ChatGPT rephrases human reviews. They employ a Transformer-based model that has been specially designed for this objective and SHAP for explainability. The issue set is on par with our objectives. This research might act as a foundation and a manual to assist us comprehend what kinds of transformer models or architectures have already been researched and how distinctive we need to be to beat existing solutions.

Regarding our topic of the identification of machine-generated text from large language models (LLMs), the work by Sadasivan et al. is directly relevant [10]. According to the experiments conducted for this paper, watermarking detectors, zero-shot classifiers, and other detectors can all

be outperformed by straightforward paraphrase. Even sophisticated, iterative paraphrase can trick detectors built to resist such assaults.

The research by Crothers et al. highlights how difficult it is to discern between material that has been created by machines and information that has been authored by people, especially with the development of strong NLG systems like ChatGPT [4]. These systems have a lot of promise, but there are also big risks of abuse. Furthermore, it acts as a guideline and a regulator for us as we design a text detection system that it's crucial for detection systems to be trustworthy.

An in-depth experiment comparing the output of five specific AI content detectors with content from three AI chatbots is shown in the Chaka paper [2]. In contrast to preceding abstractions, it adds a multilingual component by translating ChatGPT outputs into five languages before testing. Specific data from the study show the effectiveness (or lack thereof) of each detector. The failure of GPTZero to recognize translations as being created by AI was one intriguing finding.

Our approach is unique as we have fine-tuned large models and have utilized a very large and new dataset to better differentiate machine-generated text as compared to previous works both in terms of performance and time.

3. Technical Approach

We are planning to the apply three approaches to solve this:

1. **Baselines**: The SemEval already provides baselines so we will try to first replicate their approach. For the baseline we fine-tuned RoBERTa, XLM-RoBERTa and Longformer transformer for various tasks. Our ultimate goal is to outperform the baseline models in the below-mentioned approaches.

2. **Fine tuning models**: For the given tasks, in addition to implementing the baselines from scratch, we fine-tuned and experimented with various BERT based models like DistillBERT, DeBERTa-v3-large and a BERT embeddings based transformer classifier.

   (a) **DeBERTa-v3-large**, which stands for Decoding-enhanced BERT with disentangled attention, is designed to capture bidirectional context more effectively by incorporating disentangled attention mechanisms. This feature allows DeBERTa to better understand the nuances and patterns in text, making it more suitable for detecting machine-generated text from human text compared to RoBERTa. The disentangled attention in DeBERTa helps in discerning subtle differences in language usage, syntactic structures, or semantic cues that may be indicative of machine-generated text. RoBERTa, while a powerful model, does not have the same level of disentangled attention, potentially making it less effective in distinguishing between machine-generated and human-generated text.

   (b) **DistilBERT** is a smaller and faster variant of BERT, designed to have a smaller memory footprint and be more computationally efficient while maintaining competitive performance. If resource constraints such as limited computing power or memory are a consideration, DistilBERT may be a more practical choice. DistilBERT achieves computational efficiency through techniques like knowledge distillation, which involves training a smaller model (DistilBERT) to mimic the behavior of a larger model (BERT) during training.

3. **BERT Embedding based Transformer Classifier**:

   This method involves our exploration of a tailored classifier. Initially, we transform all input text embeddings using a BERT [5] model. These embeddings are subsequently fed into a customized Transformer Encoder layer-based transformer model, followed by fully-connected linear layers. Each embedding derived from BERT has a dimensionality of 768 dimensions. By harnessing this approach, we aim to enhance classification accuracy and robustness in handling diverse text inputs, leveraging the rich contextual information provided by BERT embeddings.

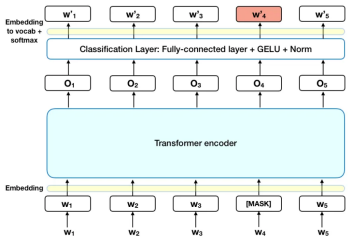


Figure 1. BERT Architecture

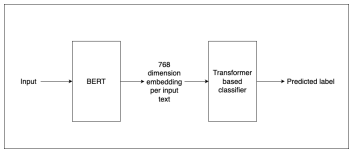


Figure 2. Custom Classifier Architecture

4. Dataset

The dataset for our project is an extension of the M4 dataset, which is described in [13] and the statistics are shown in Fig 6 . For subtask A, the monolingual dataset has 119,757 training samples and 5,000 dev samples. For the multilingual dataset, we have 172,417 training samples and 4,000 dev samples. For multilingual approximately 122,000 instances of human-machine parallel data were collected, comprising 101,000 for English, 9,000 each for Chinese, Russian, Urdu, Indonesian, and Arabic.

Additionally, over 10 million non-parallel human-written texts were included. Data consists of human-written text from diverse sources like Wikipedia, WikiHow, Reddit, arXiv, Peer-Read, Baike, and news for English, Chinese, Urdu, Russian, and Indonesian. Machine generated text involves prompts from various multilingual models like ChatGPT, Davinci003, Cohere, Dolly-v2, and BLOOMz. Tasks included writing Wikipedia articles, generating abstracts, peer reviews, answering questions, and composing news briefs.

The datasets that we have chosen are several hundred megabytes in size: the monolingual train set is of the size 331.1 MB, while the multilingual train set is 559.5 MB, which will facilitate proper batched training. Similarly, we have datasets for subtask A, subtask B and subtask C in the following format:

```
{
  "id": "identifier of the example",
  "label": "label (human text: 0, machine text: 1,)",
  "text": "text generated by a machine or written by a human",
  "model": "model that generated the data",
  "source": "source (Wikipedia, Wikihow, Peerread, Reddit, Arxiv
    )  on English or language (Arabic, Russian, Chinese,
    Indonesian, Urdu, Bulgarian, German)"
}
```

Figure 3. JSON Format for Subtask A

```
{
  "id": "identifier of the example",
  "label": "label (human: 0, chatGPT: 1, cohere: 2, davinci: 3,
    bloomz: 4, dolly: 5)",
  "text": "text generated by machine or written by human",
  "model": "model name that generated data",
  "source": "source (Wikipedia, Wikihow, Peerread, Reddit, Arxiv
    ) on English"
}
```

Figure 4. JSON Format for Subtask B

```
{
  "id": "identifier of the example",
  "label": "label (index of the word split by whitespace where
    change happens)",
  "text": "text generated by machine or written by human"
}
```

Figure 5. JSON Format for Subtask C

Source/ Domain	Language	Total Human	Parallel Data Human	Davinci003	ChatGPT	Cohere	Dolly-v2	BLOOMz	Total
Wikipedia	English	6,458,670	3,000	3,000	2,995	2,336	2,702	3,000	17,033
Reddit ELI5	English	558,669	3,000	3,000	3,000	3,000	3,000	3,000	18,000
WikiHow	English	31,102	3,000	3,000	3,000	3,000	3,000	3,000	18,000
PeerRead	English	5,798	5,798	2,344	2,344	2,344	2,344	2,344	17,518
arXiv abstract	English	2,219,423	3,000	3,000	3,000	3,000	3,000	3,000	18,000
Baike/Web QA	Chinese	113,313	3,000	3,000	3,000	–	–	–	9,000
RuATD	Russian	75,291	3,000	3,000	3,000	–	–	–	9,000
Urdu-news	Urdu	107,881	3,000	–	3,000	–	–	–	9,000
id_newspapers_2018	Indonesian	499,164	3,000	–	3,000	–	–	–	6,000
Arabic-Wikipedia	Arabic	1,209,042	3,000	–	3,000	–	–	–	6,000
True & Fake News	Bulgarian	94,000	3,000	3,000	3,000	–	–	–	9,000
Total			35,798	23,344	32,339	13,680	14,046	14,344	133,551

Figure 6. Statistics about the Dataset.

5. Experiments

5.1. Loss functions:

1. **Task A**: The primary loss function here would be binary cross-entropy loss and the error metric here will be accuracy.

2. **Task B**: The primary loss function here would be categorical cross-entropy loss and the error metric here will be accuracy.

3. **Task C**: The primary loss function here would be mean absolute error. This metric measures the absolute distance between the predicted word and the actual word where the switch between human and machine occurs. These all are the official metrics for the competition and we plan to use the same for our work as well.

5.2. Baselines Models

We split the dataset into 80-20% train-test size. Below is the list of baseline models we have used and their description.

5.2.1 RoBERTa Classfier

The RoBERTa Classifier is built upon the pre-trained language model RoBERTa, as introduced by Liu et al [7] in 2019. It has been fine-tuned specifically for the task of identifying machine-generated text. Using pre-trained classifiers has been extensively explored and applied in prior research. When RoBERTa is fine-tuned using the output of GPT-2, it demonstrates remarkable accuracy in detecting machine-generated text and showcases its capability to adapt to various decoding methods. We are using it as a baseline for Task A (monolingual) & B. Overall, we achieved test accuracy score 0.77 and 0.73 for Task A (monolingual) and Task B respectively.

5.2.2 XLM-RoBERTa

We are using XLM-RoBERTa as a baseline for Task A (multilingual). The XLM-R classifier is derived from the XLM-RoBERTa model, as introduced by Conneau [3] in 2019.

Subtask	Train	Dev
Subtask A (monolingual)	119757	5000
Subtask A (multilingual)	172417	4000
Subtask B	71027	3000
Subtask C	3649	505

Table 1. Number of train and dev examples per subtask.

Models	Task A - Monolingual (Accuracy)	Task A - Multilingual (Accuracy)	Task B (Accuracy)	Task C (MAE)
RoBERTa	0.77	-	0.73	-
XLM-RoBERTa	-	0.75	-	-
Longformer	-	-	-	3.84752
DistillBERT	0.75	0.587	0.61	6.72277
DeBERTa-v3-large	**0.83**	0.45	0.61	4.77822
Custom Transformer	0.65	0.54	0.56	-

Table 2. Performance Evaluation of Models Across Different Tasks. The upper section delineates baseline outcomes, while the lower section presents results attained by our proposed models.

XLM-RoBERTa is a cross-lingual variant of RoBERTa [8], and it is pre-trained on a multilingual dataset, allowing it to handle and comprehend text in multiple languages proficiently. The XLM-R-based detectors underwent fine-tuning, harnessing the features offered by the Transformers library14. We are using it as a baseline for Task A (multilingual). The versatility of XLM-RoBERTa extends beyond language barriers, enabling seamless comprehension across diverse linguistic landscapes. Leveraging its cross-lingual pre-training on a vast multilingual dataset, XLM-RoBERTa becomes adept at handling varied languages with proficiency. Through fine-tuning and harnessing the advanced capabilities of the Transformers library, the XLM-R-based detectors serve as a robust baseline for Task A (multilingual). Its adaptability and multilingual competence lay a solid foundation for exploring and understanding the complexities of language across different cultures and regions, offering invaluable insights into cross-lingual analysis and comprehension. Overall, we achieved test accuracy score 0.75 for Task A (multilingual).

5.2.3 Longformer

We are using Longformer as a baseline for Task C. A longformer [1] is a transformer model designed to handle lengthy documents. The longformer-base-4096 is a model similar to BERT, initiated from the RoBERTa checkpoint, and pre-trained for Masked Language Modeling (MLM) specifically for long documents. It can effectively process sequences with a length of up to 4,096. Longformer employs a blend of local attention through a sliding window and global attention. The global attention is adjustable by the user according to the specific task requirements, enabling the model to learn task-specific representations. We are using it as a baseline for Task C. Overall, we achieved Mean Absolute Error (MAE) of 3.84752 for Task C.

5.3. Proposed Models

Here is the list of models we tried for all three subtasks:

5.3.1 DistilBERT

We are running DistilBERT as our first model for tasks A, B and C. With almost 66 million parameters, the "distilBERT-base-uncased" light-weight NLP model is derived from DistilBERT [11], a simplified version of BERT. When working with lowercase text, the "uncased" feature implies that it handles uppercase and lowercase characters equally. With this design, training time and vocabulary size are reduced. When compared to the original BERT model with 110 million parameters, "distilbert-base-uncased" offers computational savings without significantly sacrificing speed. When it comes to named entity identification, text classification, and question answering, it performs remarkably well. The model is available through the Hugging Face Transformers library, which makes it easy to include into many NLP applications. Its reduced size makes it suitable for environments with constrained resources and applications that need fast inference. We are using this model for all the tasks that we have mentioned. All the results achieved using DistilBERT are being reported in Table 2.

5.3.2 DeBERTa-v3-large

The second model we ran for all three tasks A, B and C is DeBERTa-v3-large. DeBERTa V3 Large [6] stands out in

Models	Task	Learning Rate	Epochs	Batch Size	16-32 bit training (fp16)
RoBERTa	A	0.00002	3	16	32
	C	0.00005	3	32	32
XLM-RoBERTa	B	0.00002	3	16	32
Longformer	C	0.00002	3	32	32
DistillBERT	A	0.00002	3	16	32
	B	0.00002	3	16	32
	C	0.00005	3	32	32
DeBERTa-v3-large	A	0.00002	3	8	16
	B	0.00002	3	16	16
	C	0.00005	25	32	16
Custom Transformer	A	0.00001	25	32	32
	B	0.00001	25	32	32

Table 3. Model Hyperparameter Details

accuracy for both human and machine-generated text due to its highly sophisticated design and advanced training methods. Its disentangled attention [12] system dissects complex connections within the text, capturing subtle contextual cues for deeper understanding.

Argument	Value
learning_rate	2e-5
per_device_train_batch_size	8
per_device_eval_batch_size	8
num_train_epochs	2
weight_decay	0.01
gradient_accumulation_steps	2
fp16	True

Table 4. DeBERTa-v3-large Training Arguments.

The improved mask decoder precisely fills in missing elements, enhancing its grasp across various text styles. By embracing an ELECTRA-Style pre-training method and Gradient Disentangled Embedding Sharing, the model excels at deciphering intricate patterns within extensive datasets, honing a nuanced comprehension of language intricacies. With its extensive 24-layer depth, 1024 hidden size, and expanded vocabulary, DeBERTa V3 effortlessly navigates the intricacies of diverse text origins, resulting in unmatched accuracy across a wide spectrum of natural language tasks. All the results achieved using DeBERTa-v3-large are being reported in Table 2. We are aware that DeBERTa-v3-large is not suitable for multilingual settings, so, understandably, it did not perform very well.

5.3.3 BERT Embedding based Transformer Classifier

We ran the custom classifier for Tasks A and B. As mentioned in the Technical Approach section, we first fetch the embeddings of all the input texts in the subtaskA dataset for both monolingual and multilingual settings. Post that we use the TranformerEncoder based classifier with number_of_attention_heads = 8, number_of_epochs = 25, learning_rate = 1e-5 and batch_size = 32. The model performed decently well and we got accuracy scores of 0.65, 0.54 and 0.56 for SubtaskA monolingual, SubtaskA multilingual and SubtaskB respectively. All the hyperparamters involved and results achieved using the custom classifier are being reported in Table 5 and Table 6.

Hyperparameter	Value
Number of attention heads	8
Number of epochs	25
Learning rate	1e-5
Batch size	32

Table 5. Custom Classifier Hyperparameter configuration.

Subtask	Accuracy
Subtask A (monolingual)	0.65
Subtask A (multilingual)	0.54
Subtask B	0.56

Table 6. Custom Classifier Performance.

6. Comparisons & Discussions

Let us start by comparing the baseline models with our first model - DistilBERT. As illustrated in Figure 7, we manage to achieve nearly identical accuracy in Subtask A monolingual with DistillBERT, rivaling the performance of the RoBERTa baseline despite the constraint of a smaller model size. Moreover, our training time for Task A with DistillBERT is approximately 4 hours, a noteworthy improvement compared to the 10 hours required for the RoBERTa

baseline. This not only underscores the efficiency of DistillBERT in terms of model size but also highlights its advantage in training time over RoBERTa. Such insights are invaluable in real-world scenarios characterized by time limitations, computational constraints, and resource scarcity. The marginal difference of 0.02 in parity becomes negligible and diminishes in significance.

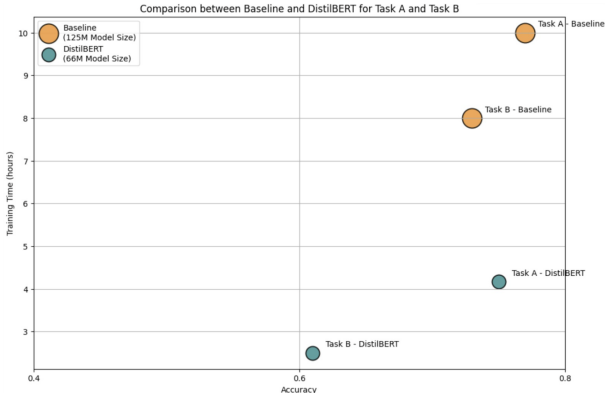
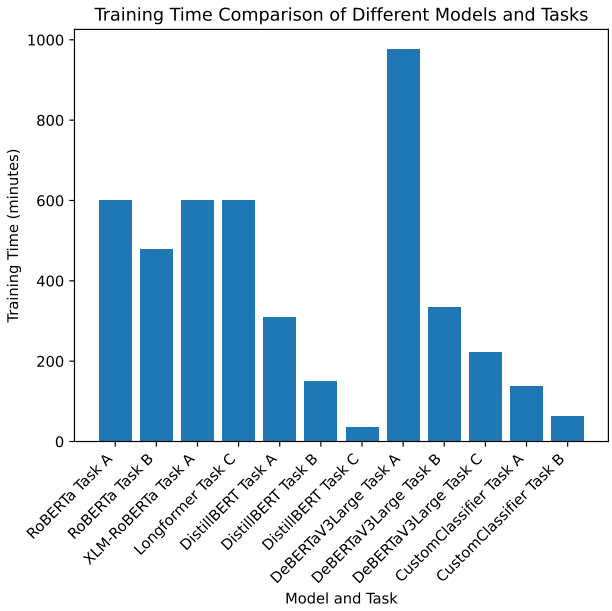


Figure 7. Task A vs Task B

In Task A - Part 2, DistilBERT shows slightly lower accuracy (0.587) compared to the baseline (0.75), but it compensates with a shorter training time of 5 hours and 10 minutes versus 10 hours. Unfortunately, there's no available information for DeBERTa on this task. In Task C, both DistilBERT and DeBERTa perform only slightly less than the baseline, with DistilBERT achieving a larger error value of 6.72277 (compared to the baseline's 3.84752) but doing so in just 35 minutes of training, as opposed to the baseline's 10 hours. DeBERTa attains a score of 4.77822, nearly reaching the baseline level while needing only 3 hours and 42 minutes for training, in contrast to the baseline's 10-hour requirement. Finally, in Task B, DistilBERT again outperforms the baseline (0.61 vs. 0.73) and achieves this with a notably shorter training time of 2 hours and 30 minutes compared to the baseline's 8 hours.

We ran the baseline model proposed by the SemEval organizers. We have used the Nvidia T4 GPU for running the models mentioned above. The results are presented in Table 2. We are able to replicate the accuracy of 77% across three runs compared to 74% reported by them for Task A monolingual dataset. It took around 10 hours to train this model. For the multilingual dataset on Task A, we are able to replicate the accuracy of 70% across three runs compared to 72%. Similarly for Task B, we are able to replicate the accuracy of 73% across three runs compared to 75% reported by them. It took around 10 hours to train this model. Lastly for Task C, we are able to replicate 3.84752 compared to 3.53 ± 0.212 reported by them. It took around 8 hours to train this model.



7. Results Analysis

In this section, we present the results of our experiments, comparing the performance of various models across multiple tasks. Our baseline models include RoBERTa, XLM-RoBERTa, and Longformer. The selected tasks encompass a range of challenges in natural language processing, including monolingual and multilingual classification (Task A), binary classification (Task B), and regression (Task C).

7.1. Task A - Monolingual Classification

Our experiments reveal compelling insights into the effectiveness of different models for monolingual classification tasks. Table 2 summarizes the accuracy scores, with DeBERTa-v3-large emerging as a standout performer, surpassing the baseline models with an accuracy of 0.83. DistillBERT also demonstrates competitive performance, while Custom Transformer lags behind with an accuracy of 0.65.

7.2. Task A - Multilingual Classification

The evaluation of models on multilingual classification tasks unveils the importance of language diversity in model selection. XLM-RoBERTa showcases its multilingual capability with an accuracy of 0.75, highlighting its effectiveness in handling diverse language inputs. DistillBERT also proves to be a robust performer in this context.

7.3. Task B - Categorical Classification

For categorical classification tasks, RoBERTa serves as a solid baseline, achieving an accuracy of 0.73. DistillBERT and DeBERTa-v3-large exhibit comparable perfor-

mance, each with an accuracy of 0.61. However, Custom Transformer falls short, yielding an accuracy of 0.56.

7.4. Task C - Regression

In the regression task, as measured by mean absolute error (MAE), Longformer stands out as the model of choice, achieving a remarkably low MAE of 3.84752. DeBERTa-v3-large and DistillBERT also exhibit competitive performance with MAEs of 4.77822 and 6.72277, respectively.

8. Implications

8.0.1 Model Architecture Impact

The superior performance of DeBERTa-v3-large across multiple tasks emphasizes the crucial role of model architecture in achieving high accuracy. This finding underscores the need for nuanced architectural choices to address task-specific challenges.

8.0.2 Multilingual Considerations

XLM-RoBERTa's success in multilingual classification tasks underscores its effectiveness in handling language diversity. This observation has important implications for applications requiring models capable of processing content in multiple languages.

8.0.3 Regression Task Suitability

Longformer's exceptional performance in the regression task suggests its suitability for tasks involving the prediction of continuous values. This finding opens avenues for exploring attention mechanisms in the context of regression-based computer vision applications.

8.0.4 Limitations of Custom Transformer

The mixed performance of Custom Transformer highlights potential limitations in its architecture for the evaluated tasks since it has only limited number of fully connected linear layers and involve less number of attention heads compared to the larger models we have used for analysis. Further analysis and refinement may be required to enhance its applicability.

9. Conclusion

In this study, we undertook a comprehensive exploration of three subtasks dedicated to discerning human-written and machine-generated text in both monolingual and multilingual scenarios. Our strategy encompassed the replication of SemEval baselines using RoBERTa, XLM-RoBERTa, and Longformer, with the aim of surpassing their performance. The integration of advanced models, such as DeBERTa-v3-large with its disentangled attention and DistilBERT for computational efficiency, highlighted the importance of nuanced attention mechanisms and practical considerations in model selection. Furthermore, by utilizing rich contextual data from BERT embeddings, our innovative BERT Embedding based Transformer Classifier demonstrated potential in improving classification robustness and accuracy. This work not only provides important insights but also opens up avenues for future research, such as optimizing the proposed classifier architecture and investigating alternative pre-trained language models, as we advance the state-of-the-art in the recognition of machine-generated text from human text. If we had access to more extensive compute resources, we should be able to explore larger models like Llama-2 in the future. Also, it would be interesting to see if this work can be generalizable to the other AI generative models in the future. All things considered, our diverse methodology offers a comprehensive perspective on text classification, opening the door for further developments in natural language processing.

10. Team Contribution:

All the team members contributed equally for the project and report analysis-related work.

A	B
Team Member	**Contributions/Responsibilities**
Aheli	Experimentation with RoBERTa and XLM-RoBERTa for task A and B
Harsh	Experimentation with DeBERTa-v3-large for task A and custom classifier for all three tasks
Aviral	Experimentation with DeBERTa-v3-large for task B & C and LongFormer for task C
Mohit	Experimentation with DistilBERT for all three tasks

Figure 8. Contribution Table

References

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 4

[2] Chaka Chaka. Detecting ai content in responses generated by chatgpt, youchat, and chatsonic: The case of five ai content detection tools. *Journal of Applied Learning and Teaching*, 6(2), 2023. 2

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 3

[4] Evan Crothers, Nathalie Japkowicz, and Herna L Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 2023. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[6] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021. 4

[7] Yinhan Liu, Myle Ott, and Naman Goyal. Jingfei du, mandar joshi, danqi chen, omer levy, mike lewis, luke zettlemoyer, and veselin stoyanov. 2019. roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 3(1), 2019. 3

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[9] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*, 2023. 1

[10] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023. 1

[11] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 4

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[13] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, et al. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*, 2023. 3