MasonTigers at SemEval-2024 Task 8: Performance Analysis of Transformer-based Models on Machine-Generated Text Detection

Sadiya Sayara Chowdhury Puspo*, Md Nishat Raihan*, Dhiman Goswami*, Al Nahian Bin Emran, Amrita Ganguly, Özlem Uzuner
George Mason University, USA
spuspo@gmu.edu

Abstract

This paper presents the *MasonTigers'* entry to the SemEval-2024 Task 8 - Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection. The task encompasses Binary Human-Written vs. Machine-Generated Text Classification (Track A), Multi-Way Machine-Generated Text Classification (Track B), and Human-Machine Mixed Text Detection (Track C). Our best performing approaches utilize mainly the ensemble of discriminator transformer models along with sentence transformer and statistical machine learning approaches in specific cases. Moreover, zero-shot prompting and fine-tuning of FLAN-T5 are used for Track A and B.

1 Introduction

In academia and beyond, machine-generated content is proliferating across news platforms, social media, forums, educational materials, and scholarly works. Breakthroughs in large language models (LLMs), like GPT-3.5 and GPT-4, facilitate the creation of fluent responses to diverse user queries. While this capability raises prospects of replacing human labor in various tasks, concerns arise about potential misuse, including the generation of deceptive misinformation (Chen and Shu, 2023) and completing student assignments, which hinders the development of essential writing skills (Jungherr, 2023).This highlights the importance of developing automated systems to detect and mitigate the potential abuse of machine-generated content, as well as distinguishing between machine-written and human-generated text. Additionally, Prior studies (ZeroGPT [1]; Mitchell et al., 2023; Bao et al., 2023) predominantly adopted a binary classification approach for machine-generated text (MGT), with a primary focus on English. However, there has been limited research addressing the amalgamation of human-written and MGT texts (Wang et al., 2024d).

In response to these limitations, SemEval-2024 introduces a shared task: Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection (Wang et al., 2024c). This task comprises three subtasks, each targeting different aspects of machine-generated text complexity.Subtask A focuses on Binary Human-Written vs. MGT Classification, involving two tracks: monolingual and multilingual. Subtask B tackles Multi-Way Machine-Generated Text Classification to identify the source of a given text. Subtask C involves detecting the transition point within a mixed text, determining where it shifts from human-written to machine-generated. The data provided for this task is an expansion of the M4 dataset (Wang et al., 2024d) and benchmark evaluation of (Wang et al., 2024b). .

In conducting these tasks, we conduct a range of experiments and observe that ensemble methods outperform individual models significantly in classification tasks, e.g., Goswami et al. (2023) Emran et al. (2024), Ganguly et al. (2024). Our weighted ensemble approaches achieve accuracies of 74%, 60% and 65% in subtask A monolingual; multilingual tracks and subtask B respectively, given that we have used different models for both tasks. In subtask C, we explore different setups, ensembling which results in Mean Absolute Error (MAE) of 60.78. For the classifications, we utilize zero-shot prompting and fine-tuning of FlanT5[2], while adhering to the restriction of no data augmentation in this task.

2 Related Works

The difficulties of separating human-written text from large language models and the significance of

*** denotes equal contribution.**
[1] www.zerogpt.com/
[2] huggingface.co/google/flan-t5-base/

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1364–1372
June 20-21, 2024 ©2024 Association for Computational Linguistics

Source	Train					Dev	
	chatGPT	**Cohere**	**Davinci**	**Dolly**	**Human**	**Bloomz**	**Human**
arxiv	3000	3000	2999	3000	15498	500	500
peerread	2344	2342	2342	2344	2357	500	500
reddit	3000	3000	3000	3000	15500	500	500
wikihow	3000	3000	3000	3000	15499	500	500
wikipedia	2995	2336	3000	2702	14497	500	500
Total				**54406**	**63351**	**2500**	**2500**

Table 1: Label Distribution of Train and Validation Data for Binary Human-Written vs. Machine-Generated Text Classification (Subtask A - Monolingual)

trustworthy methods for evaluation are highlighted by recent research (e.g. Chaka 2024, Elkhatat et al. 2023). In terms of human evaluation of MGT, Guo et al. (2023) indicates that generated texts from large language models tend to exhibit less emotional and objective content compared to human-written texts. Tang et al. (2023) suggests that distinct signals left in the machine-generated text may facilitate the identification of suitable features to differentiate between MGT and human-written texts. Whereas, Sadasivan et al. (2023) observes that detection techniques become less effective as language models improve. Moreover, Ippolito et al. (2019) advocates for the importance of using both human and automatic detectors to assess the humanness of text generation systems.

Previous work in determining MGT from human-written ones include higher order n-grams (Gallé et al., 2021), utilizing linguistic patterns (Muñoz-Ortiz et al., 2023), curvature-based criterion (Mitchell et al., 2023), tweaking with multiple variables (Dugan et al., 2023), fine-tuning transformer-based models e.g., Capobianco; Chen and Liu (2023). Very recently, Wang et al. (2024a) puts forward LLM-Detector, offering a fresh method for identifying text at both document and sentence levels by employing Instruction Tuning of LLMs. To tackle challenges of this field, several datasets have been released, e.g., MULTITuDE (Macko et al., 2023), M4 (Wang et al., 2024d). Additionally, there have been multiple shared tasks organized related to this topic (Shamardina et al., 2022a; Molla et al., Molla et al.; Kashnitsky et al., 2022. Despite several collective findings and techniques, as argued by Sadasivan et al. (2023), there remains a critical need for the creation of reliable detection methods capable of accurately distinguishing between human and machine-generated text, a requirement essential across both English and other languages.

3 Datasets

Wang et al. (2024d) collects datasets from a variety of sources, including Wikipedia (the March 2022 version), WikiHow (Koupaee and Wang, 2018), Reddit (ELI5), arXiv, PeerRead (Kang et al., 2018)(for English), and Baike (for Chinese). They employ web question answering for Chinese, news content for Urdu, Indonesian, and RuATD (Shamardina et al., 2022b) for Russian language. The method of prompting machine-generated text (MGT) has been extensively outlined in Wang et al. (2024d).

Subtask A, Binary Human-Written vs. Machine-Generated Text Classification, in the monolingual track involves a same-domain cross-generator experiment, where instances are exclusively in English and gathered from five distinct sources with two labels: 0 and 1. Human-generated texts receive a label of 0, while machine-generated texts from four different LLMs (chatGPT, Cohere, *davinci-003*, and Dolly-v2) are labeled as 1. The distribution of Train and Validation datasets, both in terms of labels and sources, along with the number of test instances, is detailed in Tables 1. During the test phase, there are 16,272 instances labeled as 0 and 18,000 instances labeled as 1.

On the other hand, Subtask A in the multilingual track entails a cross-domain same-generator experiment. Instances are sourced from nine different sources during the training phase, including four different languages, while the validation dataset comprises three different languages as indicated in Table 2. Similar to the monolingual task, human-generated texts are labeled as 0, and machine-generated texts from five different LLMs (Bloomz (Muennighoff et al., 2022), chatGPT, Cohere, *davinci-003*, and Dolly-v2) are labeled as 1. In the test phase, there are 20,238 instances labeled as 0 and 22,140 instances labeled as 1.

Source	Train						Dev		
	Bloomz	chatGPT	Cohere	Davinci	Dolly	Human	ChatGPT	Davinci	Human
arxiv	3000	3000	3000	2999	3000	15998	-	-	-
peerread	2334	2344	2342	2344	2344	2857	-	-	-
reddit	2999	3000	3000	3000	3000	16000	-	-	-
wikihow	3000	3000	3000	3000	3000	15999	-	-	-
wikipedia	2999	2995	2336	3000	2702	14997	-	-	-
Bulgarian	0	3000	0	3000	0	6000	-	-	-
Chinese	0	2970	0	2964	0	6000	-	-	-
Indonesian	0	3000	0	0	0	2995	-	-	-
Urdu	0	2899	0	0	0	3000	-	-	-
Arabic	-	-	-	-	-	-	500	0	500
German	-	-	-	-	-	-	500	0	500
Russian	-	-	-	-	-	-	500	500	1000
Total					**83571**	**83846**		**2000**	**2000**

Table 2: Label Distribution of Train and Validation Data for Binary Human-Written vs. Machine-Generated Text Classification (Subtask A - Multilingual)

Source	Train						Dev					
	Bloomz	chatGPT	Cohere	Davinci	Dolly	Human	Bloomz	chatGPT	Cohere	Davinci	Dolly	Human
arxiv	3000	3000	3000	2999	3000	2998	-	-	-	-	-	-
reddit	2999	3000	3000	3000	3000	3000	-	-	-	-	-	-
wikihow	3000	3000	3000	3000	3000	2999	-	-	-	-	-	-
wikipedia	2999	2995	2336	3000	2702	3000	-	-	-	-	-	-
peerread	-	-	-	-	-	-	500	500	500	500	500	500
Total	**11998**	**11995**	**11336**	**11999**	**11702**	**11997**	**500**	**500**	**500**	**500**	**500**	**500**

Table 3: Label Distribution of Train and Validation Data for Multi-Way Machine-Generated Text Classification (Subtask B)

Label	Test Data
Human (0)	3000
chatGPT (1)	3000
cohere (2)	3000
davinci (3)	3000
Bloomz (4)	3000
Dolly (5)	3000
Total	18000

Table 4: Label Distribution of Test Data for Multi-Way Machine-Generated Text Classification (Subtask B)

Subtask B, Multi-Way Machine-Generated Text Classification, represents another cross-domain same-generator experiment. In contrast to Subtask A, Subtask C involves six labels: 0 for human, 1 for chatGPT, 2 for Cohere, 3 for *davinci-003*, 4 for Bloomz, and 5 for Dolly. These labels correspond to instances sourced from five different sources. However, it's noteworthy that the sources for the training and validation data differ, and this distinction is outlined in Tables 3 and 4.

Subtask C, involving Human-Machine Mixed Text Detection, provides a composite text with a human-written first part followed by a machine-generated second part. The task is to discern the boundary, and labels are provided as word indices to distinguish it. The label distribution of data is shown in Table 5.

Data	Count
Train	3649
Dev	505
Test	11123

Table 5: Number of Instances for Human-Machine Mixed Text Detection (Subtask C)

4 Experimental Setup

4.1 Data Preprocessing

In the monolingual track of subtask A, we received approximately 160K instances for training and development. To preserve the text's integrity, we eliminate special characters, extra new lines, unnecessary whitespace, and hyperlinks from the data, ensuring that only the essential text remains in subtassk A (monolingual), B & C. However, in the multilingual track of subtask A, since none of our team members are familiar with the languages present in the instances, we only remove hyperlinks. We ensure that punctuation marks such as full stops, commas, and exclamation signs are retained in all instances, as they play a crucial role in this task (Tang et al., 2023).

4.2 Hyperparameters

In our experimental setup, we configure several key parameters to train our model effectively. We utilize a batch size of 16, controlling the number of training samples processed in each iteration, learning being set to 1e-5, and incorporating dropout with a rate of 0.25 to prevent overfitting by randomly dropping a fraction of units during training. Maintaining a fixed sequence length of 512 tokens ensured consistency in input data processing. For optimization, we employ the AdamW optimizer (Loshchilov and Hutter, 2017), known for its efficacy in training deep neural networks with added weight decay regularization. These experiments are conducted on a 80GB NVIDIA A100 GPU machine over the period of 24 hours, leveraging its computational power and memory capacity. By systematically adjusting these parameters, we aim to understand their influence on the model's performance, ultimately optimizing our approach for the task at hand. The adjustment of these parameters is carried out in both subtask A & B.

4.3 Models: SubTask A

In monolingual track, we employ Roberta (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and ELECTRA (Clark et al., 2020). Subsequently, we apply a weighted ensemble method, incorporating RoBERTa, DistilBERT, and ELECTRA, employing a voting strategy due to their closely comparable individual accuracies. The weights are their corresponding accuracy.

Similarly, in the multilingual track, we utilize LASER (Li and Mak, 2020), mBERT (Devlin et al., 2018), and XLMR (Goyal et al., 2021). Following that, we deploy a weighted ensembled strategy involving these models, utilizing the voting method.

4.4 Models: SubTask B

Subtask B, poses a considerable challenge, as opposed to the first two tracks where the model distinguishes between human and machine-generated text. Here, the model must differentiate among human-generated text and five distinct LLMs. For this, we leverage Roberta, ELECTRA, Deberta (He et al., 2020), and subsequently create a weighted (weights are set as acauracy) ensemble approach of these models using voting technique.

4.5 Models: SubTask C

In subtask C, we find the embedding of the training data using Term Frequency - Inverse Document Frequency (TF-IDF) (Aizawa, 2003), Positive Point-wise Mutual Information (PPMI) (Church and Hanks, 1990), and the embedding using language model RoBERTa (Liu et al., 2019). Then for each training embedding generated by these approaches, we apply Linear Regression (Groß, 2003) and ElasticNet (Zou and Hastie, 2005) separately on these embeddings and predict the first word or index of from where the machine-generated text started in a specific data instance. We selected the word that is the starting word of the closest neighboring paragraph as the predicted word index. Then we clip the predicted values to ensure the predictions range from 0 to the length of the specific data instance (rounded if necessary). In the development phase, we find the Mean Absolute Error (Chai and Draxler, 2014) of these six predictions (three each by Linear Regression and ElasticNet). Then we perform a weighted ensemble depending on the Mean Absolute Error of the six predicted results and get our ensembled MAE in the development phase. We also perform this approach on the test data and find our smallest MAE in the evaluation phase.

4.6 Prompting and Fine-Tuning LLM

Prompt: <*Text*> If this piece of text is Human Generated, answer 0 or If Machine Generated, answer 1.

Figure 1: Sample FlanT5 prompt.

For subtasks A & B, we experiment with FlanT5 zero-shot prompting, utilizing the Hugging Face Transformers[3] library, specifically the T5ForConditionalGeneration class and T5Tokenizer. Training is conducted on an NVIDIA A100 GPU with 80GB memory over 24 hours. The prompting sample for subtask A is shown in Figure 1. In subtask B, we maintain consistency in prompting by keeping the question the same as

[3] huggingface.co/docs/transformers/

Monolingual			Multilingual		
Model	**Dev**	**Test**	**Model**	**Dev**	**Test**
Baseline (RoBERTa)	0.74	0.88	**Baseline (XLM-R)**	0.72	0.81
FLAN-T5 Prompting	0.49	0.52	FLAN-T5 Prompting	0.42	0.39
FLAN-T5 Fine-tuning	0.57	0.53	FLAN-T5 Fine-tuning	0.48	0.43
RoBERTa	0.70	0.73	LASER	0.52	0.50
DistilBERT	0.69	0.70	mBERT	0.57	0.58
ELECTRA	0.78	0.71	XLMR	0.61	0.59
Ensemble (Wt. accuracy)	0.79	0.74	**Ensemble (Wt. accuracy)**	0.63	0.60

Table 6: Accuracy of Binary Human-Written vs. Machine-Generated Text Classification (Subtask A)

labeling the human-generated text as "1", while prompting the machine-generated texts from various Language Model Models (LLMs) as categories "2" through "6."

We also finetune a t5-small model over 2 epochs, setting the learning rate to 0.001 and the batch size to 4. We employ a full finetuning (FFT) approach without the utilization of any quantization method like LoRa (Hu et al., 2021) or QLoRA (Dettmers et al., 2023). Due to the adoption of an FFT approach and the sheer size of the dataset, we do not experiment with a wide set of hyper-parameters. We empirically choose a few combinations and report the best results.

5 Results

Subtask A and B are evaluated based on Accuracy, as specified by (Wang et al., 2024c), while Subtask C employs Mean Absolute Error (MAE) as the evaluation metric [4].

In the monolingual track of Subtask A, ELECTRA demonstrates superior accuracy (0.78) compared to RoBERTa (0.70) and DistilBERT (0.69) during the development phase. Consequently, the weighted ensemble of these three models achieves an accuracy of 0.79 in our development submission, surpassing the baseline RoBERTa model. Upon publishing test labels, a comparison with the test label results reveals accuracies detailed in Table 6, with the ensemble model achieving an accuracy of 0.74, while the baseline accuracy increases to 0.88, differing by 0.14 compared to the development phase. In the multilingual track, XLM-R outperforms LASER and mBERT with an accuracy of 0.61. Ensembling these models achieves accuracies of 0.63 in the development phase and 0.60 in the test phase, whereas the baseline accuracies are 0.72 and 0.81, respectively. Both zero-shot prompting and fine-tuning FlanT5 demonstrate less than satisfactory performance, yielding accuracies of 0.53 and 0.43 in the monolingual and multilingual tracks, respectively.

Model	Dev	Test
Baseline (RoBERTa)	0.75	0.75
FLAN-T5 Prompting	0.54	0.48
FLAN-T5 Fine-tuning	0.57	0.54
RoBERTa	0.72	0.56
ELECTRA	0.73	0.59
DeBERTa	0.77	0.64
Ensemble (Wt. accuracy)	0.79	0.65

Table 7: Accuracy of Multi-Way Machine-Generated Text Classification (Subtask B)

Within subtask B, DeBERTa outperforms RoBERTa and ELECTRA, achieving superior performance with an accuracy of 0.77. Ensembling these models yields accuracies of 0.79 and 0.65 in both the development and test phases, whereas baseline RoBERTa gives 0.75 in both phases. Similar to subtask A, fine-tuning and prompting FLAN T5 exhibit suboptimal results in both phases shown in Table 7.

In subtask C, various methods are considered, and it is found that RoBERTa with Elastic Net achieved the minimum Mean Absolute Error (33.28). Table 8 highlights that Elastic Net outperforms Linear Regression in terms of lower MAE during both the development and test phases. To enhance predictive performance, we employ a weighted ensemble of development phase MAE of six combinations, resulting in MAE values of 31.71 and 60.78 during the development and test phases, respectively. However, the baseline (longformer) model gives MAE of 3.53 ± 0.21 and 21.54.

[4] https://github.com/mbzuai-nlp/SemEval2024-task8

Model	Dev	Test
Baseline (Longformer)	$\simeq$ 3.53	21.54
TF-IDF + LR	44.15	71.23
PPMI + LR	41.93	68.41
RoBERTa + LR	37.52	65.82
TF-IDF + EN	38.36	67.09
PPMI + EN	35.67	63.36
RoBERTa + EN	33.28	62.34
Wt. (dev. MAE) Ensemble	31.71	60.78

Table 8: Mean Absolute Error(MAE) value of Human-Machine Mixed Text Detection (Subtask C) (LR = Linear Regression, EN = ElasticNet)

6 Error Analysis

In the monolingual track of Subtask A, the final model demonstrates proficiency in accurately identifying machine-generated text. Nonetheless, there is a notable presence of false positives, indicating instances where the model incorrectly identifies human-written texts as machine-generated. Despite this, the model effectively detects machine-generated text without omission. Similarly, in the multilingual track of Subtask A, the ultimate model excels in accurately distinguishing machine-generated text. However, false positives are prevalent, indicating numerous cases where human-written texts are inaccurately classified as machine-generated. Additionally, the model encounters instances where it fails to predict machine-generated texts.

In Subtask B, the model excels in accurately predicting chatGPT-generated texts. However, its performance declines notably for davinci-generated text, often misclassifying it as chatGPT generated. Additionally, the model's accuracy is lower for Dolly-generated and human-written texts, indicating a discrepancy in handling machine-generated versus human-written content.

For subtask C, MAE is higher due to the presence of outliers because the dev MAE was significantly lower than the test MAE. To handle this issue, it is essential to address the preprocessing of data, handling outliers, selecting appropriate features, optimizing model complexity, improving data quality, and ensuring model stability through proper tuning and evaluation procedures. This can be the future scope of research in this specific domain.

For a clearer understanding, refer to the visual evaluations in Figure 2, 3, 4 of Appendix.

7 Conclusion

In our investigation of SemEval-2024 Task 8, we applied a diverse set of methodologies, encompassing statistical machine learning techniques, transformer-based models, sentence transformers, and FLAN T5. Subtask A involved binary classification, where the monolingual track focused on cross-generator scenarios within the same domain, and the multilingual track addressed cross-domain scenarios within the same generators. Subtask B dealt with multi-label classification, requiring the discrimination of human-generated text from five distinct language models. Subtask C centered on Human-Machine Mixed Text Detection, employing TF-IDF, PPMI, and RoBERTa with Linear Regression and ElasticNet for prediction. The outcomes of three subtasks highlighted the efficacy of ensemble methods, showcasing specific models excelling in each subtask. Additionally, we explored the applicability of zero-shot prompting and fine-tuning FLAN-T5 for Tracks A and B.

In summary, our approach harnessed a blend of transformer models, machine learning methodologies, and ensemble strategies to tackle the complexities presented by SemEval-2024 Task 8. The paper underscores the imperative need for robust detection methods to effectively navigate the growing prevalence of machine-generated content.

Limitations

The task involved extensive datasets in each phase of all subtasks, leading to prolonged execution times and increased GPU usage. Additionally, the texts themselves were lengthy. Moreover, the prohibition of additional data augmentation added to the complexity of the task. The nuanced distinction between human-written and machine-generated text, which can sometimes be challenging for humans to discern, poses an even greater difficulty for models attempting to learn this differentiation.

Acknowledgements

We express our gratitude to the organizers for orchestrating this task and to the individuals who diligently annotated datasets across various languages. Your dedication has played a crucial role in the triumph of this undertaking.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Marc Capobianco. *Supervised Machine Generated Text Detection Using LLM Encoders In Various Data Resource Scenarios*. Ph.D. thesis, WORCESTER POLYTECHNIC INSTITUTE.

Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*.

Chaka Chaka. 2024. Reviewing the performance of ai detection tools in differentiating between ai-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching*.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Zheng Chen and Huming Liu. 2023. Stadee: Statistics-based deep detection of machine generated text. In *Proceedings of ICIC*, pages 732–743. Springer.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

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

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of AAAI*.

Ahmed M Elkhatat, Khaled Elsaid, and Saeed Almeer. 2023. Evaluating the efficacy of ai content detection tools in differentiating between human and ai-generated text. *International Journal for Educational Integrity*.

Al Nahian Bin Emran, Amrita Ganguly, Sadiya Sayara Chowdhury Puspo, Dhiman Goswami, and Md Nishat Raihan. 2024. Masonperplexity at climateactivism 2024: Integrating advanced ensemble techniques and data augmentation for climate activism stance and hate event identification. *arXiv preprint arXiv:2402.01976*.

Matthias Gallé, Jos Rozen, Germán Kruszewski, and Hady Elsahar. 2021. Unsupervised and distributional detection of machine-generated text.

Amrita Ganguly, Al Nahian Bin Emran, Sadiya Sayara Chowdhury Puspo, Md Nishat Raihan, Dhiman Goswami, and Marcos Zampieri. 2024. Masonperplexity at multimodal hate speech event detection 2024: Hate speech and target detection using transformer ensembles. *arXiv preprint arXiv:2402.01967*.

Dhiman Goswami, Md Nishat Raihan, Sadiya Sayara Chowdhury Puspo, and Marcos Zampieri. 2023. nlpbdpatriots at blp-2023 task 2: A transfer learning approach to bangla sentiment analysis. In *Proceedings of BLP (EMNLP)*.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of RepL4NLP*.

Jürgen Groß. 2003. *Linear regression*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*.

Andreas Jungherr. 2023. Using chatgpt and other large language model (llm) applications for academic paper assignments.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*.

Yury Kashnitsky, Drahomira Herrmannova, Anita de Waard, Georgios Tsatsaronis, Catriona Fennell, and Cyril Labbé. 2022. Overview of the dagpap22 shared task on detecting automatically generated scientific papers. In *Proceedings of SDP*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Wei Li and Brian Mak. 2020. Transformer based multilingual document embedding model. *arXiv preprint arXiv:2008.08567*.

Yinhan Liu, Myle Ott, 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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, et al. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. *arXiv preprint arXiv:2310.13606*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Diego Molla, Haolan Zhan, Xuanli He, and Qiongkai Xu. 2023. Overview of the 2023 alta shared task: Discriminate between human-written and machine-generated text. *Training*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and llm-generated text. *arXiv preprint arXiv:2308.09067*.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected?

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

Tatiana Shamardina, Vladislav Mikhailov, Daniil Cherniasnkii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022a. Findings of the the ruatd shared task 2022 on artificial text detection in russian. In *Computational Linguistics and Intellectual Technologies*.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Chernianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022b. Findings of the the ruatd shared task 2022 on artificial text detection in russian. *arXiv preprint arXiv:2206.01583*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts.

Rongsheng Wang, Haoming Chen, Ruizhe Zhou, Han Ma, Yaofei Duan, Yanlan Kang, Songhua Yang, Baoyu Fan, and Tao Tan. 2024a. Llm-detector: Improving ai-generated chinese text detection with open-source llm instruction tuning. *arXiv preprint arXiv:2402.01158*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, et al. 2024b. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. *arXiv preprint arXiv:2402.11175*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, jinyan su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2041–2063, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024d. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of EACL*.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

A Appendix

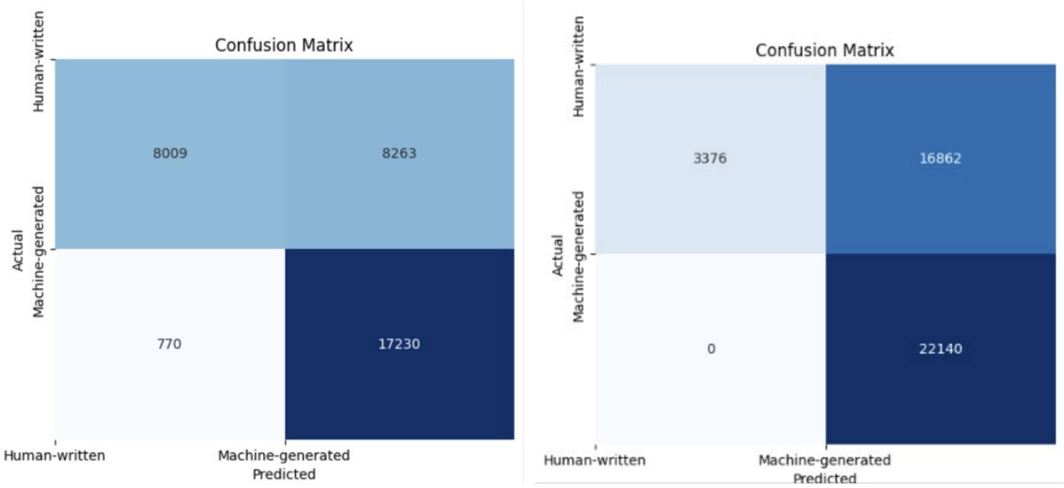

Figure 2: Confusion Matrix (Binary Human-Written vs. Machine-Generated Text Classification : Monolingual (Left), Multilingual (Right))

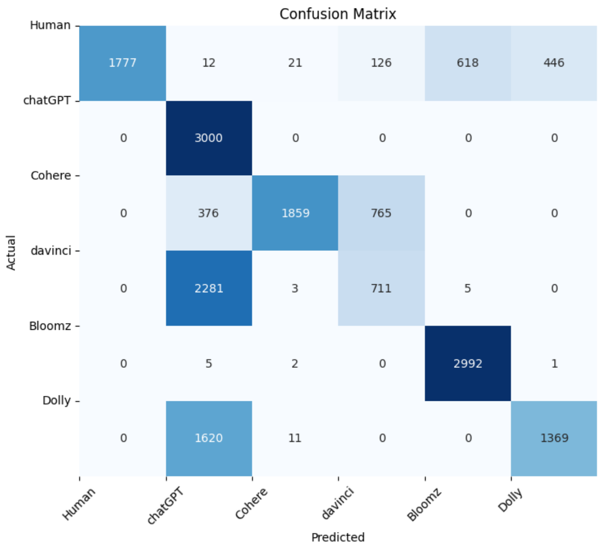

Figure 3: Confusion Matrix (Multi-Way Machine-Generated Text Classification)

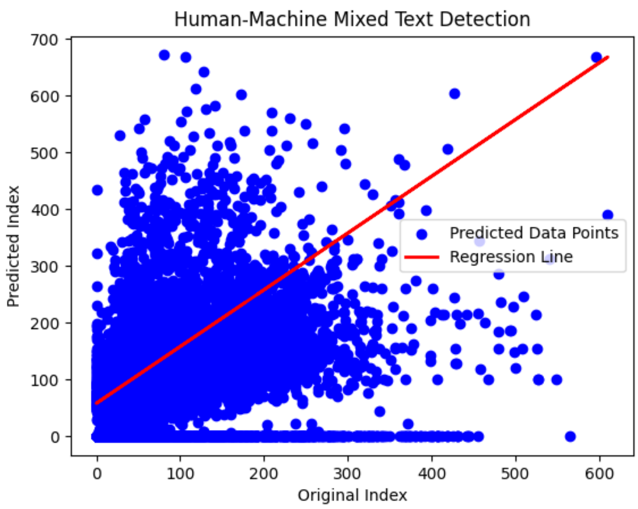

Figure 4: Regression (Human-Machine Mixed Text Detection)