Team MLab at SemEval-2024 Task 8: Analyzing Encoder Embeddings for Detecting LLM-generated Text

Kevin Li†
Stanford University
kevinli7@stanford.edu
Kenan Hasanaliyev†
Stanford University
kenanhas@stanford.edu
Sally Zhu
Stanford University
salzhu@stanford.edu
George Altshuler
Stanford University
gwa@stanford.edu
Alden Eberts
Stanford University
ajeberts@stanford.edu
Eric Chen
Stanford Univeristy
ericc27@stanford.edu
Kate Wang
Stanford University
kyw1923@stanford.edu
Emily Xia
Stanford University
emxia18@stanford.edu
Eli Browne
Stanford University
ebrowne@stanford.edu
Ian Chen
Stanford University
ianychen@stanford.edu
Umut Eren
Stanford University
umuteren@stanford.edu

Abstract

This paper explores solutions to the challenges posed by the widespread use of LLMs, particularly in the context of identifying human-written versus machine-generated text. Focusing on Subtask B of SemEval 2024 Task 8, we compare the performance of RoBERTa and DeBERTa models. Subtask B involved identifying not only human or machine text but also the specific LLM responsible for generating text, where our DeBERTa model outperformed the RoBERTa baseline by over 10% in leaderboard accuracy. The results highlight the rapidly growing capabilities of LLMs and importance of keeping up with the latest advancements. Additionally, our paper presents visualizations using PCA and t-SNE that showcase the DeBERTa model's ability to cluster different LLM outputs effectively. These findings contribute to understanding and improving AI methods for detecting machine-generated text, allowing us to build more robust and traceable AI systems in the language ecosystem.

1 Introduction

We live in a society that currently relies heavily on the use of LLMs (Large Language Models), which has followed from the explosive popularity of ChatGPT when it was released in late 2022. Now, with the introduction of GPT-4 and other more powerful LLMs, it has become increasingly important for us to have the ability to distinguish human-written text from machine-generated text. The fluency of recent models, paired with their tendency to hallucinate, has given rise to a very natural concern that there could be both accidental and intentional "bad actors" seeking to spread false information. Research has indicated that about one in every five jobs has over half of its tasks incorporated into LLMs, and that statistic is positively correlated with the barrier to entry (Eloundou et al., 2023). Consequently, these models have the necessary training data to spit out an immense number of plausibly correct but actually incorrect texts, which would be extremely detrimental because humans historically have been unable to distinguish between them beyond the level of random guessing. Such results are supported with recent work aiming to distinguish human-written sentences from AI-generated ones, with the AuTexTification study additionally demonstrating that cross-domain AI-generated text detection from Bloomz or GPT is more difficult in non-English languages (Sarvazyan et al., 2023). Furthermore, in efforts to facilitate unbiased dataset generation for related studies, the framework of TextMachina was created, and it contains crucial post-processing abilities like removing disclosure patterns and truncation (Sarvazyan et al., 2024). SemEval-2024's Task 8 (Wang et al., 2024) attempts to provide a working solution

† Indicates equal contribution amongst authors

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1463–1467

to the above problems by using standalone, self-operational means to classify whether a given text was authentically written by a human or artificially generated by a machine, in hopes of eventually building toward a foolproof method of detecting misinformation.

In subtask A, we were tasked with creating a binary classification model to determine if a given text was human-written or machine-generated. Within this subtask, there are two different tracks: one for monolingual (only English) and another for multilingual sources (something about which track we did or if we ended up doing both). Out of curiosity, we used base DeBERTa with default hyperparameters as our submission for this subtask, but it only performed 5% worse than the RoBERTa baseline. This specific subtask is important because LLMs are becoming more powerful and easily accessible, so there is a larger potential for misuse. This classification task would help catch the people who are misusing this technology to harm society.

In Subtask B, we trained a neural network to identify not only whether a given text was human-written or machine-written, but also to identify which large language model was responsible for generating that text. These language models include ChatGPT, Cohere, Dolly, and more. This task is important for many of the same reasons as subtask A; as LLMs are becoming more capable and accessible, being able to distinguish between human and model is crucial. Furthermore, being able to distinguish between different models allows for better enforcement of AI-safety laws and accountability.

2 Methods

2.1 RoBERTa

The baseline performances provided for both subtasks A and B revolved around HuggingFace's RoBERTa model. RoBERTA was developed to enhance the usability of post-BERT models, and incorporated a variety of techniques including longer training times, larger batches, more data, the elimination of the next sentence prediction objective, longer training sequences, and dynamic modification of the masking pattern (Liu et al., 2019). For the monolingual component of subtask A, roberta-base was used for a baseline of about 0.88, and for the multilingual component, xlm-roberta was used (to account for the various other languages) for a baseline of about 0.81. For subtask B, roberta-base was used again for a baseline of about 0.75.

2.2 DeBERTa

DeBERTa was designed as an upgrade to BERT and RoBERTa with the addition of disentangled attention and an enhanced mask decoder, and furthermore, its fine-tuning included adversarial training (He et al., 2021). As a result, we used the deberta-base model from HuggingFace with the assumption that it would outperform RoBERTa, and our best model did. Regarding hyperparameter tuning, we wanted to fine-tune the deberta-base model on the given dataset, so we looped through learning rates of 1e-5, 5e-5, and 1e-4, batch sizes of 4 and 8, epoch counts of 2 and 3, and weight decay constants of 1e-3, 5e-3, and 1e-2. By truncating the input length to a constant 1024 tokens, we established that the optimal hyperparameters (at least from what we tested) that yielded the highest accuracy were a learning rate of 1e-5, a batch size of 4, an epoch count of 3, and a weight decay constant of 1e-2.

2.3 Model interpretibility

To further analyze the inner working of the DeBERTa model, we analyzed our trained model's pooled outputs that encode the input sentence as whole prior to the logits. We used two dimensionality reductions algorithms, namely PCA and t-SNE for 2-D projection. PCA operates by finding orthogonal directions with the highest variance and projecting to the subspace spanned by the orthogonal directions. The t-SNE algorithm (van der Maaten and Hinton, 2008) works by preserving pairwise similarities in the data to generate related clusters. Our t-SNE projections were computed with a perplexity of 35 and iteration count of 300. Both algorithms were run on the 18,000 sentences in the test data for subtask B.

3 Results

The final DeBERTa model had the following results on the validation set, with a weighted average of 0.98633 precision, 0.98599 recall, and 0.98599 F1-score.

In comparison, our RoBERTa model performed worse, with each F1-score being lower than the corresponding F1-score for the DeBERTa model.

This model had weighted scores of 0.97979 precision, 0.97909 recall, and 0.97909 F1-score. Although these were high, they were each still a

Label	Source	Precision	Recall	F1-Score
0	Human	0.99916	0.99375	0.99645
1	ChatGPT	0.94944	0.99417	0.97129
2	Cohere	0.98735	0.99824	0.99276
3	Davinci	0.98912	0.94708	0.96765
4	Bloomz	1.0	0.99833	0.99917
5	Dolly	0.99311	0.98610	0.98606

Table 1: DeBERTa results on the Subtask B validation set

Label	Source	Precision	Recall	F1-Score
0	Human	0.99916	0.98792	0.99351
1	ChatGPT	0.93008	0.99250	0.96027
2	Cohere	0.97631	1.0	0.98801
3	Davinci	0.98367	0.92875	0.95542
4	Bloomz	0.99791	0.99667	0.99729
5	Dolly	0.99170	0.96966	0.98055

Table 2: RoBERTa results on the Subtask B validation set

bit lower than the corresponding metrics for the DeBERTa model.

We compared our model's predictions for the test set with the labels provided. The total accuracy was 0.8266666667. Below is the confusion matrix in table form for the predicted labels versus actual labels, with the labels corresponding to their respective sources. For example, the 439 entry has predicted label 2 and actual label 3, meaning that there were 439 predictions for the Cohere source that were actually from the Davinci source.

	Actual					
Pred	0	1	2	3	4	5
0	2050	3	9	237	541	151
1	0	2823	6	171	0	0
2	11	208	2342	439	0	11
3	27	624	3	2334	12	27
4	0	1	0	1	2997	1
5	77	187	34	626	541	4612

Table 3: Confusion matrix for Subtask B labels

The results on the test set are summarized in the table below.

Additionally, we analyzed the pooled outputs of the trained DeBERTa model on subtask B's data. To visualize the outputs in 2-D, PCA and t-SNE projection techniques were applied on the 768-D pooled outputs. The data points were colored by their corresponding text source (either human or LLM model) in Figures 1 and 2 on the test set.

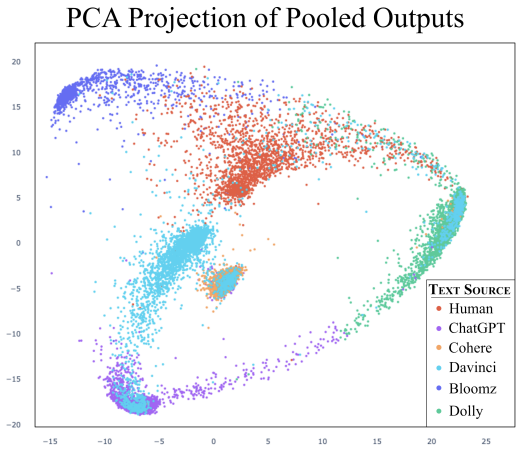

Figure 1: PCA projection of pooled outputs on the subtask B test data

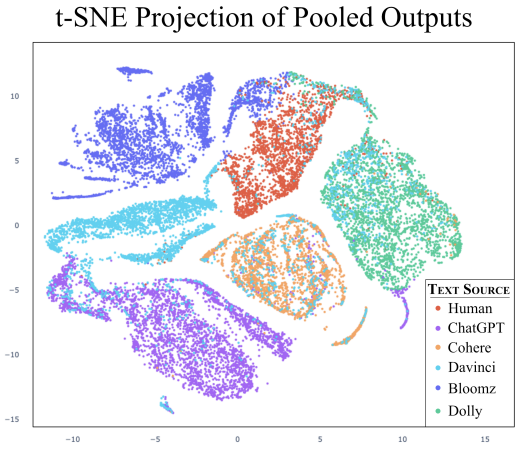

Figure 2: t-SNE projection of pooled outputs on the subtask B test data

Label	Source	Precision	Recall	F1-Score
0	Human	0.95	0.69	0.80
1	ChatGPT	0.73	0.94	0.82
2	Cohere	0.98	0.78	0.87
3	Davinci	0.65	0.78	0.71
4	Bloomz	0.84	1.00	0.91
5	Dolly	0.93	0.78	0.85

Table 4: DeBERTa test results on Subtask B

4 Discussion

4.1 Validation and Test Set Results

For both the RoBERTa statistics and the DeBERTa statistics in their validation results, the top two entries for precision and F1-score were Bloomz and Human in some order, which indicates a greater degree of identifiability from these sources. Bloomz was trained on multilingual tasks and fine-tuned on English prompts (Muennighoff et al., 2023), and many online texts are written by humans who are at least bilingual, so there could be a mannerism of text generation tied to multilingualism that makes it easier to pinpoint and distinguish these sources from the others.

Another observation is that Davinci performed reasonably well across the board for the validation sets, but had abysmal scores for the DeBERTa test set. Considering that DeBERTa is a more recent model, it is entirely possible that it has more difficulty with older data, which may explain why Davinci did as poorly as it did. However, besides Davinci and Bloomz being outliers on either end of the spectrum for F1-score, the rest of the values fell within a generally stable range, indicating that DeBERTa had a balanced evaluation of texts.

Additionally, interestingly enough, ChatGPT ranks last or near last in precision and F1-score in all the tables, but makes up for that with its high recall values. This could mean that the text was detected to be AI-generated with relative ease, but was then often misclassified as being from another AI source. Given that GPT-4 has greatly enhanced abilities compared to its predecessor, swapping out ChatGPT for GPT-4 could yield radically different results (for a potential future direction).

4.2 Visualizations of Pooled Outputs

It is clear from both the PCA and t-SNE visualizations that the DeBERTa model is successfully able to distinguish between different LLMs and human output in distinct clusters. Of note, however, are the blue points corresponding to Davinci text located in clusters of different colors. This phenomenon follows from recent research and shows that human writing tasks can still be quite susceptible to LLM influence due to their positive association with exposure (Eloundou et al., 2023). We speculate that Davinci being one of the earliest models influenced the training data of the other models that came on later, causing them to write similarly to Davinci. This supports our earlier hypothesis from the raw results, but seemingly contradicts findings that Davinci exhibits fewer confusions and is thus easily distinguishable from other models (Sarvazyan et al., 2023). One possible explanation for this is that our visualizations used parameters that clearly confined the other sources to their regions; it is entirely possible that a different configuration of parameters would yield a graph that displays an obvious Davinci scatter area while having a jumble of colors elsewhere for the other models.

5 Conclusion

For SemEval 2024 Task 8, Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection, Team MLab submitted models for subtasks A and B. Specifically for subtask B, we used a base DeBERTa model and significantly outperformed the provided baseline set by a base RoBERTa model, with our model's final accuracy coming out to 0.827. In analyzing the precision, recall, and F1-score statistics, we discovered trends in the recorded values that seem to indicate that the method of training models, as well as the timeline of their training, have profound effects on the detectability of machine-generated text. Finally, by creating and interpreting PCA and t-SNE graphs, we present visual evidence that DeBERTa's internal reasoning groups various LLM results in separate clusters, even though Davinci acted as an exception with its colored points scattered in the general vicinity of other models. Therefore, visualizing AI thought processes can provide us with useful insights regarding how we can understand and improve the language ecosystem that they share with us.

References

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

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.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

Areg Mikael Sarvazyan, José Ángel González, and Marc Franco-Salvador. 2024. Textmachina: Seamless generation of machine-generated text datasets.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

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. 2024. Semeval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico, Mexico.