Feature-Level Ensemble Learning for Robust Synthetic Text Detection with DeBERTaV3 and XLM-RoBERTa

Saman Sarker Joy and **Tanusree Das Aishi**
Department of Computer Science and Engineering
BRAC University, 66 Mohakhali, Dhaka-1212, Bangladesh
{saman.sarker.joy, tanusree.das.aishi}@g.bracu.ac.bd

Abstract

As large language models, or LLMs, continue to advance in recent years, they require the development of a potent system to detect whether a text was created by a human or an LLM in order to prevent the unethical use of LLMs. To address this challenge, ALTA Shared Task 2023 introduced a task to build an automatic detection system that can discriminate between human-authored and synthetic text generated by LLMs. In this paper, we present our participation in this task where we proposed a feature-level ensemble of two transformer models namely DeBERTaV3 and XLM-RoBERTa to come up with a robust system. The given dataset consisted of textual data with two labels where the task was binary classification. Experimental results show that our proposed method achieved competitive performance among the participants. We believe this solution would make an impact and provide a feasible solution for detection of synthetic text detection.

1 Introduction

In recent years, the remarkable advancements in Large Language Models (LLMs) have showcased an unprecedented revolution in the field of Natural Language Processing (Raiaan et al., 2023). These models e.g. GPT-X and T5, demonstrate the ability to generate text that closely resembles content created by humans. However, there is a risk of abuse that makes this a double-edged sword and raises moral questions. The spread of synthetic text produced by LLMs carries the risk of spreading false information (Bian et al., 2023), interfering in elections (Schneier, 2023), and jeopardize the credibility of scientific knowledge (Birhane et al., 2023).

In this context, the ALTA Shared Task 2023[1] introduced a task where researchers have to develop automated detection systems with the capacity to discriminate between human-written text and text generated by Large Language Models (LLMs). The aim of this task is to mitigate the unethical application of LLMs and promote their conscientious and responsible utilization in various domains.

The dataset provided by the ALTA Shared Task 2023 is used in building and assessing the automatic text detection systems. The task is fundamentally a binary classification problem. Each text is labeled as either 0 (AI-generated) or 1 (human-generated). The text samples are derived from diverse domains, including law and medicine, and span a spectrum of LLMs. The efficacy of the models will be assessed based on their accuracy and their resilience in detecting synthetic text. This paper presents our approach to this task, where we at first performed some data analysis on the dataset. Then using those analysis, we have proposed a feature-level ensemble model that utilizes the strengths of two state-of-the-art transformer models: DeBERTaV3 and XLM-RoBERTa. We believe that our proposed method, refined through rigorous experimentation, has achieved competitive and robust performance, positioning it as a promising solution among the participating models.

Text	Label
I am asking you this because the fans of the band are completely devoted. They experience the days leading up to the concert very intensely.	0
A Reston man has been charged with abduction after police say he dragged a woman from the sidewalk and tried to remove her clothes.	1

Table 1: Example of ALTA Shared Task 2023. Here, the two labels are 0 (for AI-generated) or a 1 (for human-generated).

[1] https://www.alta.asn.au/events/sharedtask2023/description.html

Text	Label	ID	Language Found
Rektor på Gammel Hellerup Gymnasium, Jørgen Rasmussen, ønsker ikke at udtale sig om sagen.	1	36	German
En réalité, la superteam qui semble se profiler est composée de :	1	5628	French
E le parole “programma di aggiustamento strutturale”, “ristrutturazione” e “default” si possono benissimo tradurre in genocidio sociale.	1	15541	Italian

Table 2: Examples of text in the train set containing other languages.

2 Dataset Description

The dataset statistics are summarized in Table 3. Each text in the dataset is tagged with either a 0 (for AI-generated) or a 1 (for human-generated). The training dataset is evenly balanced, ensuring an equal number of both categories.

In our analysis, we found that text length in the training set typically range from 10 to 50 and never exceed 200 which can be seen in Figure 1. Also, we came across that some words of the texts are in languages other than English, e.g. German, Afrikaans, Romanian, French, etc., in about 124 rows of the train dataset. Some examples are shown in Table 2. This indicates a bit of multilingual content needs to be considered as well.

Category	Data
Train	18000
Validation	2000
Test	2000

Table 3: Dataset Splits.

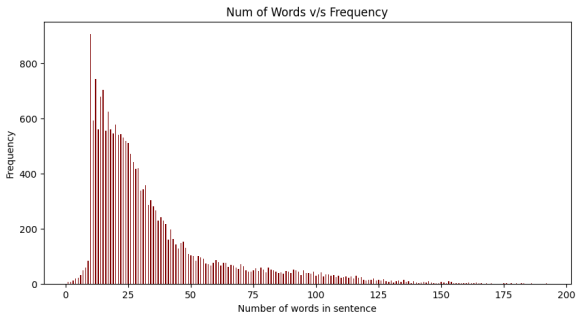

Figure 1: Frequency of all the text in the training set. It shows that all the text lengths are less than 200.

3 System Overview

3.1 Transformer Model

Transformers (Vaswani et al., 2017) are widely used in NLP tasks because they excel at various tasks. They offer high performance, scalability, and flexibility, making them a popular choice for many applications.

3.1.1 DeBERTaV3

DeBERTaV3 (He et al., 2023) is a new pretrained language model that improves upon the original DeBERTa (He et al., 2021) model. It does so by using a pretraining task called replaced token detection (RTD) instead of the traditional mask language modeling (MLM) task, which is more sample-efficient.

3.1.2 XLM-RoBERTa

XLM-RoBERTa (Conneau et al., 2019), a large-scale multilingual language model based on Facebook’s RoBERTa (Liu et al., 2019). XLM-RoBERTa undergoes pretraining on an extensive 2.5TB dataset of filtered CommonCrawl data.

3.2 Training Strategies

3.2.1 Ensamble Learning

Our model utilizes feature-level ensemble learning by independently extracting valuable information from two pre-trained models, DeBERTaV3 and XLM-RoBERTa. We extract essential details from each of these models and then effectively combine them to enhance our model’s performance. This approach is known as feature-level ensemble learning.

3.2.2 Model Architecture

We used the base versions (12 layers) of both DeBERTaV3 and XLM-RoBERTa. We made this

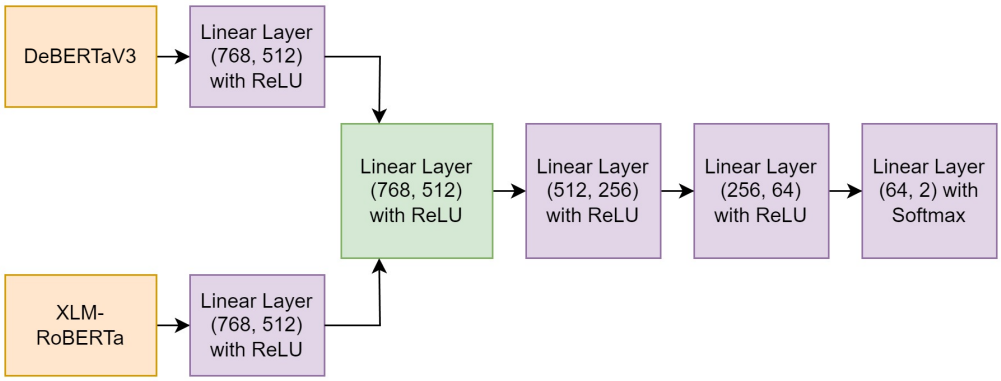

Figure 2: DeBERTaV3 and XLM-RoBERTa concatenation of the last layer + MLP Architecture.

choice due to computation limitations, as these base versions are more computationally efficient while still delivering effective results.

In this setup, the last layer of the DeBERTaV3 and XLM-RoBERTa models is used as the input. These representations are then passed through additional linear layers and fine-tuned on the specific task or dataset during the training phase. This allows the model to adapt to the task while benefiting from the pretrained language representation capabilities of DeBERTaV3 and XLM-RoBERTa. From Figure 2, we can see an overview of our model.

3.2.3 Data Augmentation

We had a validation set for which we initially did not have the correct labels. However, when we trained our initial model, we achieved a high level of accuracy in the validation set. So, we decided to include the validation data along with the predicted labels it generated into our original training dataset. This adjustment improved our models' performance.

4 Experiments and Evaluation

4.1 Experimental Settings

For hyperparameters, we have taken the number of epochs for training as 20, the learning rate is 1e-5, maximum length is 200, batch size of 8, the loss function is Cross Entropy Loss and the optimizer is AdamW (Loshchilov and Hutter, 2017).

4.2 Evaluation Metric

The ALTA Shared Task 2023 organizers employed a standard evaluation metric accuracy to evaluate the participants' system. They calculated the accuracy score using scikit-learn's (Pedregosa et al., 2011) accuracy_score package.

4.3 Results and Analysis

From Table 4, the highest accuracy in this competition was achieved by "OD-21" securing the top position with a score of 0.9910.

Our team, "SamNLP" submitted which achieved an accuracy of 0.9820, securing the 4th position in the original contest. This initial model achieved validation dataset accuracy of 0.9930, signifying that 1986 out of 2000 samples were accurately classified. Consequently, we integrated this high-performing dataset into our training data through augmentation. The model we later developed with the validation data added to the training set, was not initially submitted during the contest but was submitted after the contest had concluded. The rank that is listed 2nd is not the original rank it achieved during the contest but rather represents the rank it would have attained if it had been submitted as part of the competition.

A noteworthy observation is the marginal differences in accuracy among the top-performing teams. The variations in accuracy between the top-ranking teams are quite low, suggesting that the competition was highly competitive and challenging.

5 Conclusion

In conclusion, the growing capabilities of Large Language Models (LLMs) have brought both opportunities and challenges in the field of Natural Language Processing. The rise of synthetic text generated by LLMs has raised ethical concerns, including the spread of misinformation and potential misuse in various domains. To address this, the ALTA Shared Task 2023 was introduced. In this paper, we presented our approach to this task, where we focused on building a feature-level ensemble

Team Name	Accuracy	Position
SamNLP (Ours)**	0.9820	4th**
SamNLP (Ours with validation data added)*	0.9855	2nd*
Competitive performance of top-ranked methods		
OD-21	**0.9910**	1st
DetectorBuilder	0.9845	2nd
AAST-NLP	0.9835	3rd
Organizers	0.9765	5th
VDetect	0.9715	6th
cantnlp	0.9675	7th

Table 4: Comparative performance of our proposed method along with top-performing participants' method. The double asterisk (**) represents the actual position for the test dataset, while the single asterisk (*) denotes the model and accuracy achieved in the test dataset after the conclusion of the contest.

model using two state-of-the-art transformer models. We conducted a comprehensive analysis of the dataset, which revealed the need to handle multilingual content. Our approach leveraged feature-level ensemble learning, utilizing the strengths of both models, and included data augmentation to enhance performance. While we secured the 4th position in the original contest, the inclusion of validation data improved our model's accuracy, bringing it to the 2nd position when submitted after the contest's conclusion. Notably, the top-performing teams in the competition exhibited marginal differences in accuracy, emphasizing the high level of competitiveness in the task. We believe that our proposed method provides a promising solution for the detection of synthetic text, contributing to the responsible and conscientious use of LLMs in various applications. As LLMs continue to evolve, robust detection systems like the one presented in this paper become increasingly important to address the ethical challenges associated with AI-generated text.

References

Ning Bian, Hongyu Lin, Peilin Liu, Yaojie Lu, Chunkang Zhang, Ben He, Xianpei Han, and Le Sun. 2023. Influence of external information on large language models mirrors social cognitive patterns.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nat. Rev. Phys.*, 5(5):277–280.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Mohaimenul Raiaan, Md. Saddam Hossain, Kaniz Fatema, Nur Fahad, Sadman Sakib, Most. Marufatul Jannat Mim Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2023. A review on large language models: Architectures, applications, taxonomies, open issues and challenges.

Bruce Schneier. 2023. Ai disinformation is a threat to elections learning to spot russian, chinese and iranian meddling in other countries can help the us prepare for 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.