Team MGTD4ADL at SemEval-2024 Task 8: Leveraging (Sentence) Transformer Models with Contrastive Learning for Identifying Machine-Generated Text

Huixin Chen[1], Jan Büssing[2], David Rügamer[2,3], Ercong Nie[†1,3]
[1] Center for Information and Language Processing (CIS), LMU Munich,
[2] Institute for Statistics, LMU Munich,
[3] Munich Center for Machine Learning (MCML)
{chen.huixin, jan.buessing}@campus.lmu.de
david.ruegamer@stat.uni-muenchen.de
nie@cis.lmu.de

Abstract

This paper outlines our approach to SemEval-2024 Task 8 (Subtask B), which focuses on discerning machine-generated text from human-written content, while also identifying the text sources, i.e., from which Large Language Model (LLM) the target text is generated. Our detection system uses Transformer-based techniques and incorporates various pre-trained language models (PLMs), which are tools that help understand and process language, including sentence transformer models. Additionally, we incorporate Contrastive Learning (CL) into the classifier to improve the detecting capabilities and employ Data Augmentation methods. Ultimately, our system achieves a peak accuracy of 76.96% on the test set of the competition, configured using a sentence transformer model integrated with CL methodology.

1 Introduction

The emergence of sophisticated Large Language Models (LLMs) has significantly blurred the lines between human-written and machine-generated texts, prompting an urgent need for systems capable of accurately distinguishing between the diverse sources.

In response, our team participated in SemEval-2024 Task 8: Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection, as defined by Wang et al. (2024). This task aims to identify the origin of texts across various languages and domains, addressing critical concerns around the misuse of LLMs. We focused on Subtask B, which involves classifying English texts by their generative sources. This task adopted a fine-grained label set, for distinguishing not only between human-written and machine-generated texts, but also among texts generated by different machines. Our system leveraged Transformer-based pre-trained language models (PLMs) as well as its variant, Sentence Transformer models (Reimers and Gurevych, 2019). By applying Contrastive Learning (CL) approaches, which aimed at enhancing model robustness and generalization to our system, our best approach yielded a modest improvement over the baseline on the test set, achieving an accuracy of 76.96% compared to the baseline's 74.61%, and ranking 20th in the competition. The code for our system, detailed further in this paper, is made available at: https://github.com/banjuessing/adl_semeval24_mgtd.

2 Background and Related Work

The introduction of the M4 dataset by Wang et al. (2023) offers a comprehensive landscape for evaluating detection techniques across various generators, domains, and languages. The research done on the M4 dataset underscores the difficulties in generalizing detection across different domains and generators, highlighting the limitations of current approaches.

Data for Subtask B of SemEval-2024 Task 8, focusing on the detection of human-written against machine-generated texts from multiple generators across monolingual (English) contexts, is derived from the original M4 dataset. The dataset comprises 71,027 training and 3,000 development/test samples, distributed across multiple sources — Wikipedia, Reddit, arXiv, and wikiHow — with the testing data focused on the out-of-domain PeerRead domain. The task demands the identification of text origins, whether human or machine-generated by models. This underscores the necessity for systems that are adept at handling multi-class and out-of-domain classification challenges. In response to these challenges, our approach builds upon the insights from prior work. Abdalla et al. (2023), for instance, applied linguistic- and transformer-based method to detecting the author-

[†] Corresponding author.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1711–1718

ship of text. We also considered the methods used in the M4 dataset paper to compare with.

3 System Overview

Having outlined the urgency and relevance of distinguishing machine-generated text, we now describe our Transformer-based approach to tackle this issue. Our system tackles machine-generated text detection by carefully selecting a suite of transformer models (RoBERTa$_{\text{BASE}}$, RoBERTa$_{\text{LARGE}}$ (Liu et al., 2019), GPT-2 Small (Radford et al., 2019), XLNet-Base (Yang et al., 2019)), sentence transformer models (all-mpnet-base-v1[1], all-mpnet-base-v2[2], all-roberta-large-v1[3]), and integrating two different Contrastive Learning techniques, namely Supervised Contrastive Learning (SCL) (Gunel et al., 2020; Khosla et al., 2020) and Dual Contrastive Learning (DualCL) (Chen et al., 2022), alongside data augmentation strategies, inspired by (Bhattacharjee et al., 2023). Initially, we conducted hyperparameter tuning across both transformer and sentence transformer models in their base forms with trivial cross-entropy (CE) loss to identify optimal configurations. Subsequently, we refined our model selection to GPT-2 Small, RoBERTa$_{\text{BASE}}$, RoBERTa$_{\text{LARGE}}$, and all-roberta-large-v1, based on performance metrics on the enriched test set, which is described in section 4.1, further experimenting with combination of contrastive learning technique variants to enhance detection accuracy. Our approach culminates in the additional application of data augmentations aiming to improve robustness and generalizability.

3.1 Transformers

We began our exploration with a diverse set of transformer models: RoBERTa$_{\text{BASE}}$, RoBERTa$_{\text{LARGE}}$, GPT-2 Small and XLNet-Base, distinguished by unique architectural designs and pre-training objectives. RoBERTa, as an encoder model enhanced with optimized training approach dynamic masking strategy on longer sequences, offers robustness and depth in understanding context. GPT-2 with its generative capabilities and autoregressive training objective, provides insights into the sequence prediction dynamics often employed by text generation models. XLNet, incorporating permutation-based training, may capture bidirectional context and outperform traditional unidirectional models in understanding complex sentence structures. The different characteristics of those models grant us a comparative edge in detecting generated content. We conducted extensive hyperparameter tuning to identify the configurations that yield optimal performance on the classification task, which was crucial for ensuring that each model was leveraged to its fullest potential.

3.2 Sentence Transformers

In parallel, we evaluated three sentence transformer models: all-mpnet-base-v1, all-mpnet-base-v2, and all-roberta-large-v1. The decision to incorporate sentence transformers alongside traditional transformers was driven by their further pre-training on sentence pairs for generating semantically rich embeddings (Reimers and Gurevych, 2019), which potentially offers a more nuanced understanding of the essence of entire sentences. The selection of the three variants was informed by their pre-training paradigms and underlying architectures, which may influence their performance on text classification tasks. The all-mpnet-base models with a relatively smaller model size of 420 MB and hidden dimension of 768, derived from the MPNet (Song et al., 2020), are notable for their optimized permuted language modeling pre-training upon XLNet. The distinction between v1 and v2 primarily lies in the maximum sequence length, with v1 having 512 tokens and v2 having 384 tokens. The all-roberta-large-v1 model, on the other hand, is built upon the RoBERTa architecture with a larger model size of 1360 MB, a larger hidden dimension of 1024 but a smaller context window size of 256 tokens. Similar to the transformer models, hyperparameter tuning was performed to fine-tune these models for our specific task, ensuring that the models' configurations were optimized.

3.3 Contrastive Learning

Based on the initial evaluations, we narrowed our focus to GPT-2 Small, RoBERTa$_{\text{BASE}}$, RoBERTa$_{\text{LARGE}}$, and all-roberta-large-v1. These models were subjected to further experiments to test the efficacy of Contrastive Learning methods in enhancing their performances.

Driven by the training objectives of the sentence transformers and the intuition that in the embedding space, examples from the same source tend to

[1]https://huggingface.co/sentence-transformers/all-mpnet-base-v1

[2]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[3]https://huggingface.co/sentence-transformers/all-roberta-large-v1

be grouped together, while examples from different generators or human could be potentially pushed apart to be distinguished, we integrated SCL loss and DualCL loss with our selected models. Both loss functions utilize Contrastive Learning in the supervised setting. Following Gunel et al. (2020), the SCL loss directly takes the samples from the same class as positive samples and the samples from different classes as negative samples, while the DualCL loss simultaneously learns from the features of input samples $\mathcal{L}_z$ and the parameters of classifiers $\mathcal{L}_\theta$ in the same space with label-aware data augmentation (Chen et al., 2022). The overall loss that we used to optimise the models is then one of the two following combinations of two losses, where the λ adjusts the balance between the primary loss function and the contrastive loss component:

$$\mathcal{L}_{overall}^{SCL} = (1-\lambda)\mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{SCL}} \quad (1)$$

$$\mathcal{L}_{overall}^{DualCL} = (1-\lambda)\mathcal{L}_{\text{CE}} + \frac{1}{2}\lambda\mathcal{L}_{\text{DualCL}} \quad (2)$$

where $\mathcal{L}_{\text{DualCL}} = \mathcal{L}_z + \mathcal{L}_\theta$.

Each model was trained and evaluated using one of these Contrastive Learning methods, in addition to the traditional CE loss, to compare their effectiveness systematically. Hyperparameter tuning was again employed for each combination of model and loss to ensure optimal settings.

3.4 Hyperparameter Optimization

For hyperparameter optimization (HPO) within our detection system, we employed a grid search strategy. Specifically, when training our models using the conventional CE loss, our tuning focused solely on optimizing the learning rate of the optimizer. Conversely, in scenarios where models were trained with the incorporation of SCL loss or DualCL loss, we extended our tuning efforts to include both the learning rate and the λ value.

3.5 Data Augmentation

Finally, we investigated the role of data augmentation in further enhancing the models' ability to discern machine-generated text. Selecting the top-performing model configurations from the previous steps, we applied various data augmentation techniques using `nlpaug` library[4] (Ma, 2019), including synonym replacement and random word swap to enrich the training dataset. This step aimed to introduce variability and complexity to the training process, testing the hypothesis that augmented data could lead to better generalization and robustness.

[4]https://github.com/makcedward/nlpaug

4 Experimental Setup

4.1 Data

The dataset for SemEval-2024 Task 8 encompasses a broad spectrum of text generators, encompassing both human-authored and machine-generated sources. The machine generated texts include outputs from advanced LLMs: BLOOMz, ChatGPT, Cohere, Davinci-003, and Dolly-v2. The data features diverse domains, including arXiv, WikiHow, Wikipedia, Reddit and PeerRead. This composition challenges us to distinguish human-written text from machine-generated content and further identify the specific LLM responsible.

The dataset was strategically split by organizers to promote an out-of-domain testing scenario, with the test set solely containing PeerRead texts absent from training data, comprising only 500 samples evenly distributed across each generator. This limitation led us to enrich our test dataset by incorporating all available samples from the original M4 dataset specific to the PeerRead domain, thereby aiming for a comprehensive analysis within our experimental framework. Utilizing the full PeerRead dataset provided us access to 14,566 data points, significantly enhancing our ability to conduct a deeper exploration of text detection capabilities. The distribution of data points across each model/source is as follows:

Model/Generator	Number of Samples
BLOOMz	2,334
ChatGPT	2,344
Cohere	2,342
Davinci-003	2,344
Dolly-v2	2,344
Human	2,858

Table 1: Distribution of data points across models/sources for the Peerread domain in our enriched test dataset.

To address the requirements of our experimental setup, we partitioned the original training dataset, as provided by the organizers, into two subsets: 90% for training and 10% for validation, where the labels and source domains of the samples are evenly distributed. This division was consistently applied

across all experiments to maintain uniformity in the evaluation process. Additionally, the enriched test dataset, as previously described, was employed as the test dataset for all experimental validations.

4.2 Hyperparameters

Under the experimental setup, a consistent approach was adopted for hyperparameter selection across all models to ensure comparability of the results. We ultilized AdamW (Loshchilov and Hutter, 2017) optimizer with a default weight decay of 0.01 for training each model across all experiments. Training was conducted with mixed precision for 20 epochs, incorporating an early stopping mechanism triggered by 3 consecutive epochs of loss increase. For the hyperparameter tuning of all transformer models and sentence transformer models using CE loss, a grid search methodology was implemented. The learning rate parameters explored were {1e-5, 2e-5, 5e-5}, with the exception of the RoBERTa$_{\text{LARGE}}$ model, for which a range of {1e-6, 2e-6, 5e-6} was tested. When integrating either SCL loss or DualCL loss, We explored the learning rate combined with λ values of {0.02, 0.1, 0.2} to adjust the influence of contrastive loss. Our decisions of selecting best performed models based on the accuracy of each model's performance on our enriched test dataset.

5 Results and Discussions

In our analysis of the performance of four transformer models as our baseline on the task of detecting machine-generated text, distinct variations in accuracy underscore the impact of model design and size on effectiveness, as shown in Table 2. The GPT-2 Small model, achieving the highest accuracy at 73.25%, outperformed both XLNet-Base and RoBERTa models. This superior performance could be attributed to GPT-2's architecture, primarily designed as a decoder model for generating text, which may inherently provide it with a nuanced capability to distinguish between human and machine-generated texts. When comparing models within the same family, RoBERTa$_{\text{BASE}}$'s performance surpasses that of RoBERTa$_{\text{LARGE}}$. This observation suggests that increasing model size, and thereby complexity, does not necessarily translate to better performance in detecting machine-generated text and a smaller model might be more effective than its larger counterpart. This could be due to the diminishing returns of model capacity expansion in this specific task.

Model	Accuracy
XLNet-Base	64.22
GPT-2 Small	**73.25**
RoBERTa$_{\text{BASE}}$	67.31
RoBERTa$_{\text{LARGE}}$	64.29

Table 2: Accuracy of the transformer models on the enriched test set. The results are reported as the best performance among each model's hyperparameter configurations.

In our evaluation of three sentence transformer models, we observed distinct performance outcomes that offer insights into the influence of model architecture and input sequence length on accuracy, as shown in Table 3. Specifically, the all-mpnet-base-v1 and all-mpnet-base-v2, which share the same foundational model and architectural parameters including model size and hidden dimension, demonstrated only a marginal difference in accuracy (61.36% for v1 and 60.73% for v2). This slight discrepancy in performance, despite v1's capability to process longer input sequences than v2, suggests that an extended context window does not inherently guarantee superior detection efficacy in our task. Conversely, the all-roberta-large-v1 model, characterized by its robust architecture and a higher hidden dimension of 1024, although with a reduced context window size, markedly outperformed the aforementioned models, achieving an accuracy of 69.96%. This outcome underscores the observation that a larger context window, contrary to expectations, may not be as critical for enhancing machine-generated text detection as previously assumed.

Model	Accuracy
all-mpnet-base-v1	61.36
all-mpnet-base-v2	60.73
all-roberta-large-v1	**69.96**

Table 3: Accuracy of the sentence transformer models on the enriched test set. The results are reported as the best performance among each model's hyperparameter configurations.

Our explorations with selected best models from previous experiments further led to insightful observations regarding the performance of incorporating contrastive learning methods, as shown in Table 4. For GPT-2 Small model, both the CL losses corrupted the performance, indicating the

alignments of CL losses may not suitable for a decoder model. For the RoBERTa$_{BASE}$ model, integrating CL methodologies yielded results comparable to those obtained using traditional CE loss with a slight underperformance. Similarly, the RoBERTa$_{LARGE}$ model, when augmented with CL methods, demonstrated only a marginal improvement under 2% over the conventional CE loss approach. Conversely, the all-roberta-large-v1 sentence transformer model showed a strong contrast in performance when leveraging two contrastive learning losses. The model variant with additional SCL loss markedly outperformed the accuracy achieved with standard CE loss, resulting in the best model variant across all our experiments. However, incorporating DualCL loss resulted in substantially poorer performance compared to the baseline, hinting at potential mismatches between the DualCL objective and the sentence transformer model for the task-specific requirements. Upon comparing the overall performances, the all-roberta-large-v1 model outperformed remarkably both the RoBERTa$_{BASE}$ and RoBERTa$_{LARGE}$ models, indicating that the adaptation and specialization of sentence transformers significantly contribute to discerning the subtle intricacies of machine-generated texts with the SCL loss further enhancing this ability.

Model	Loss	Accuracy (%)
GPT-2 Small	CE	73.25
	CE+SCL	72.53
	CE+DualCL	58.15
RoBERTa$_{BASE}$	CE	67.31
	CE+SCL	66.85
	CE+DualCL	66.64
RoBERTa$_{LARGE}$	CE	64.29
	CE+SCL	64.94
	CE+DualCL	65.94
all-roberta-large-v1	CE	69.96
	CE+SCL	**74.60**
	CE+DualCL	53.16

Table 4: Accuracy of the selected best performed models with various loss functions on the enriched test set. The results are reported as the best performance among each combination's hyperparameter configurations.

Incorporating data augmentation techniques, as detailed in section 3.5, to further train the GPT-2 Small and all-roberta-large-v1 model, which demonstrated top-2 performances in previous experiments, resulted in a significant decrease in performance, details shown in Table 5 in Appendix A.1. This decline was observed across both configurations of utilizing CE loss and the combination of CE loss and SCL loss, despite their initially high accuracy on the enriched test set. A potential reason for this downturn could be the introduction of noise or irrelevant variations through data augmentation, which may have led to the models' reduced ability to generalize from the augmented data, ultimately detracting from its capability to accurately distinguish machine-generated texts.

As we analyse model performance dynamics, an intriguing pattern of overfitting emerged among some of the top-performing model configurations. Upon testing earlier checkpoints of these models against the enriched test set, it was observed that certain pre-final checkpoints exhibited superior performance compared to the final models, which had achieved the highest validation accuracy. This phenomenon suggests that models slightly earlier in their training phase, before reaching peak validation accuracy, may generalize better to unseen data when detecting the machine generated texts. We report the detailed performance dynamics in Table 6 in Appendix A.2.

As we observe our best model's performance on the enriched test dataset for each generator, we find that the model demonstrates a robust ability to accurately identify texts generated by Dolly-v2, BLOOMz, ChatGPT, and Cohere, indicating a strong alignment with the characteristics prevalent in the outputs from these sources. However, it encounters significant challenges when attempting to classify texts originating from human authors and the Davinci-003 model. We report the detailed confusion matrix in Appendix A.3. This insight points to the need for further model refinement and training to bridge the gap in detection capabilities across some certain text origins.

6 Conclusion

In conclusion, our paper presents a comprehensive approach to SemEval-2024 Task 8 (Subtask B), focusing on the detection of machine-generated text and its attribution to specific Large Language Models (LLMs). Leveraging Transformer-based methods, pre-trained language models (PLMs), Contrastive Learning (CL), and Data Augmentation techniques, we have developed a robust detection system achieving a peak accuracy of 74.69%. Our findings underscore the effectiveness of integrat-

ing CL into the classification process and highlight the strength of leveraging diverse PLMs for improved performance in discerning between human and machine-generated text.

Acknowledgments

We thank Ercong Nie for his valuable guidance and support in our participation in the SemEval-2024 Task 8: Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection as part of Applied Deep Learning Course at the Ludwig-Maximilians-Universität München organized by Prof. David Rügamer.

References

Mohamed Hesham Ibrahim Abdalla, Simon Malberg, Daryna Dementieva, Edoardo Mosca, and Georg Groh. 2023. A benchmark dataset to distinguish human-written and machine-generated scientific papers. *Information*, 14(10).

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. Conda: Contrastive domain adaptation for ai-generated text detection.

Qianben Chen, Richong Zhang, Yaowei Zheng, and Yongyi Mao. 2022. Dual contrastive learning: Text classification via label-aware data augmentation. *CoRR*, abs/2201.08702.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *CoRR*, abs/2011.01403.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *CoRR*, abs/2004.11362.

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.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297.

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

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, 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 City, Mexico.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

A Appendix

A.1 Models with Data Augmentation

Table 5 shows the detailed result of experiments implemented with Data Augmentation methods.

A.2 Model Performances Dynamics

Table 6 shows the training dynamics.

A.3 Confusion Matrix of the Best Performed Model

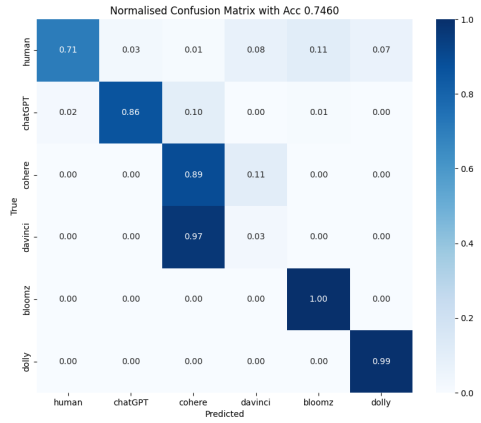

Figure 1: Confusion matrix of the our best model's (all-roberta-large-v1 with CE+SCL loss) performance on the enriched test dataset described in section 4.1 with texts only in Peerread domain.

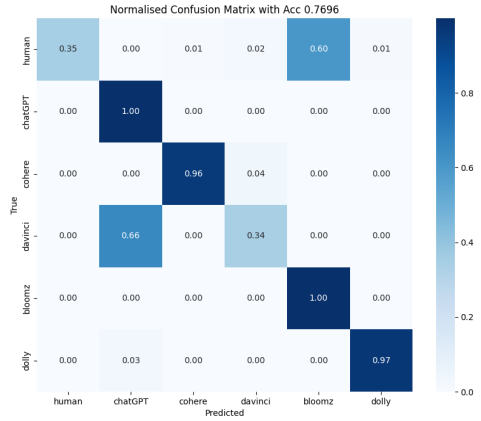

Figure 2: Confusion matrix of the our best model's (all-roberta-large-v1 with CE+SCL loss) performance on the test dataset that provided by the organizers.

Model	Loss	Augmentation	Accuracy (%)
GPT-2 Small	CE	No	73.25
		Yes	56.82
all-roberta-large-v1	CE+SCL	No	74.60
		Yes	58.81

Table 5: Accuracy of the best-performed GPT-2 Small and all-roberta-large-v1 model with various loss functions and data augmentation on the enriched test set. The results are reported as the best performance among each combination's hyperparameter configurations.

Save Point (epoch)	1	2	3	4	5	6	7	8	9	10
$\text{RoBERTa}_{\text{BASE}}$	65.34	66.04	-	58.51	**66.42**	65.47	-	-	63.50	
all-roberta-large-v1	66.15	69.57	70.93	73.05	-	-	**74.60**	-	-	70.84
GPT-2 Small	69.93	-	71.25	-	-	**72.53**				

Table 6: Accuracy(%) of the three selected model configurations' performances across different epochs on the enriched test set. We select $\text{RoBERTa}_{\text{BASE}}$ with DualCL(λ=0.02), all-roberta-large-v1 with SCL(λ=0.2) and GPT-2 Small with SCL(λ=0.2) to observe the performance dynamics, because among the best performed model configurations they have long enough converge processes. The dashes in the table indicate no model checkpoint is saved in that epoch due to no increase in the validation accuracy. Saved model checkpoints in the later epochs have higher validation accuracy.