KoBERT-Based Multi-Class Text Classification Applied to *Act on Protective Action Guidelines Against Radiation in the Natural Environment*

Hyunwoo Lee[a, b], Jongeun Kim[a, c], Minjung Kim[a], Sujin Kim[a, b], Insu Chang[b], Seungkyu Lee[b], Yoonsun Chung[a*]
[a]*Department of Nuclear Engineering, Hanyang university, 222, Wangsimni-ro, Seongdong-gu, Seoul, Korea*
[b]*Korea Atomic Energy Research Institute, 111 Daedeok-Daero, Yuseong-gu, Daejeon, Korea*
[c]*Proton Therapy Center, National Cancer Center, 323, Ilsan-ro, Ilsandong-gu, Goyang-si, Gyeonggi-do, Korea*
[*]*Corresponding author: ychung@hanyang.ac.kr*

Keywords : KoBERT, Text classification, Domain Adaptive Pre-Training, Environmental radiation

1. Introduction

The rapid development of Transformer-based large language models (LLMs) has led to significant transformations across both industry and academic fields. However, it is recognized that LLMs trained on general-purpose text data often exhibit suboptimal performance in domain-specific contexts. A representative approach to address this limitation is fine-tuning. Following pre-training and fine-tuning, the performance of language models significantly improves across various tasks, including question answering (Q&A), named entity recognition (NER), summarization, and text classification [1].

In this study, we utilized KoBERT, a BERT-based Korean language model. Furthermore, we pre-trained it on domain-specific texts related to the *Act on Protective Action Guidelines Against Radiation in the Natural Environment*. Since this field is highly specialized and narrowly defined, it was expected that a fine-tuned domain-specific language model would demonstrate performance comparable to that of a commercial LLM. Subsequently, the pre-trained KoBERT was fine-tuned to classify input sequences according to the corresponding clauses of the same act. Finally, the performance of the proposed domain-specific model was compared with that of commercial LLM to assess whether a specialized language model could achieve performance comparable to general-purpose commercial systems.

2. Materials and Methods

2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a foundational language model that uses a multi-head attention mechanism to analyze contextual relationships among words [2]. This enables it to understand the meaning of a word based on its surrounding context. KoBERT, developed for the Korean language, is a lighter version of BERT, with a reduced vocabulary (30,002→8,002) and fewer parameters (110M→92M) [3].

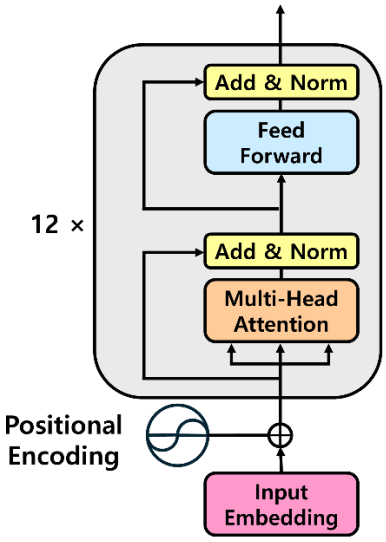

Figure 1. Structure of BERT

2.2 Dataset and Vocabulary

Dataset for pre-training was compiled from four sources: reports and yearbooks on natural environment radiation, internet news articles, and research paper abstracts. The final preprocessed corpus was composed of 5.51 M tokens. To enhance the model's domain knowledge, python module named SentencePiece was used to extract 2,000 tokens from corpus [4]. After removing tokens that overlapped with KoBERT's base vocabulary, 342 essential tokens selected by domain experts were added to a customized tokenizer.

2.3 Pre-training and Fine-tuning

Domain-Adaptive Pre-Training (DAPT) was performed on our KoBERT model using the domain-

specific corpus. The goal of the subsequent fine-tuning task was to classify sentences from the *Act on Protective Action Guidelines Against Radiation in the Natural Environment* into 31 specific clauses or a "Not related" category. A dataset of 16,000 labeled sentences was split into a 60% training set, 20% validation set, and 20% test set.

Severe class imbalance was observed in the dataset, with the majority class making up 82.1% of the data and the smallest class 0.03%. To mitigate bias from this imbalance, we conducted two parallel experiments: one using cross-entropy loss and another using weighted cross-entropy loss, which assigns weights inversely proportional to class frequency. The model was trained for 10 epochs with a fixed learning rate of 1e-05.

2.4 Metrics

Model performance was evaluated using total accuracy, macro precision, macro recall, and macro F1 score. These scores except total accuracy are calculated as the arithmetic mean of the precision, recall and F1 scores for each class, providing a balanced evaluation that is not skewed by the dominant classes. Formula for total accuracy, macro precision, macro recall and macro F1 score are presented below.

$$Total\ accuracy = \frac{Number\ of\ correct\ answers}{Number\ of\ samples\ in\ test\ set}$$

$$Precision = \sum_{i=1}^{N} \frac{P_i}{N}, P_i = precision\ of\ class\ i$$

$$Recall = \sum_{i=1}^{N} \frac{R_i}{N}, R_i = recall\ of\ class\ i$$

$$macro\ F1\ score = \sum_{i=1}^{N} \left(\frac{2 \times P_i \times R_i}{P_i + R_i}\right) / N$$

3. Results and Discussion

3.1 Fine-tuning

Figure 1 illustrates the performance improvement of the four models during fine-tuning. As shown, the models employing cross-entropy as the loss function achieved their peak performance (with the lowest loss) at epoch 6. After that, overfitting became evident. In contrast, both the training and validation losses of the models utilizing weighted cross-entropy as the loss function continued to decrease throughout the fine-tuning process.

3.2 Overall Performance Analysis

Performance of four fine-tuned KoBERT models—varying by DAPT application and loss function (cross-entropy or weighted cross-entropy)—was compared with GPT-5-mini. All models showed a significant discrepancy between total accuracy (0.82-0.88) and the macro F1 score (0.12-0.37), indicating a performance imbalance likely caused by severe class imbalance in the dataset.

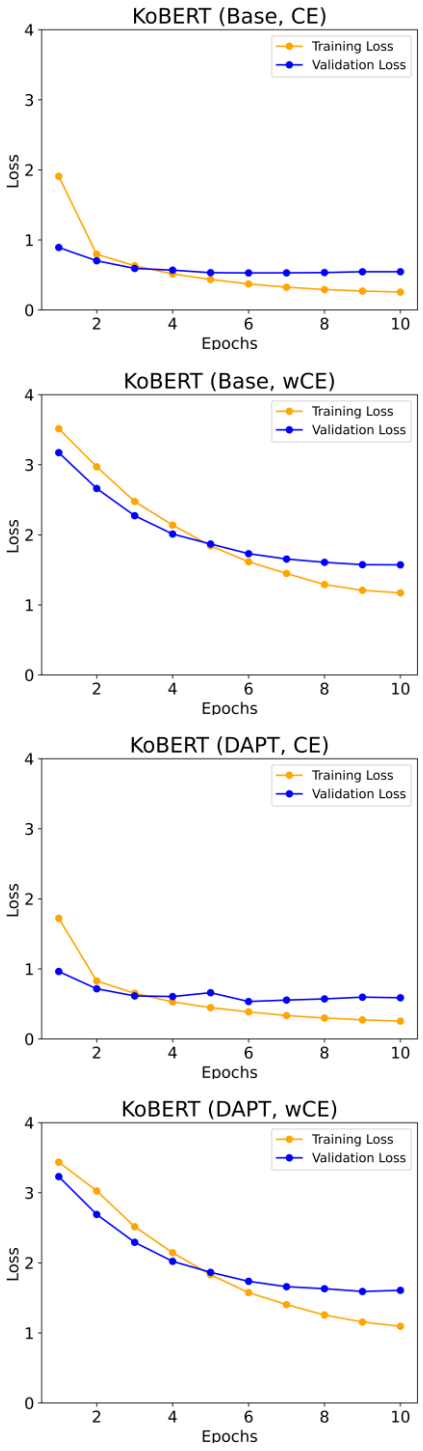

Figure 2. Changes in training and validation losses throughout fine-tuning of models.

3.3 Impact of Loss Function

The weighted cross-entropy loss function proved highly effective in mitigating the performance gap

between major and minor classes. For KoBERT (Base) and KoBERT (DAPT), switching to this loss function resulted in substantial increases in macro precision, recall, and F1 scores. However, it also led to a slight decrease in total accuracy, as it reduced the model's ability to correctly classify the majority "Not related" class, which constitutes many hard negative samples.

Table I: Precision, Recall, macro F1 score of models

Model	Overall accuracy	Precision	Recall	Macro F1
KoBERT				
Base, CE*	0.88	0.21	0.17	0.18
Base, wCE**	0.83	0.35	0.41	0.35
DAPT, CE*	0.88	0.15	0.18	0.16
DAPT, wCE**	0.82	0.36	0.45	0.37
GPT-5-mini	0.82	0.23	0.09	0.12

*Cross-Entropy
**weighted Cross-Entropy

3.4 Impact of DAPT

The positive effect of DAPT was observed only when the weighted cross-entropy loss function was used. KoBERT (DAPT, wCE) showed an improvement in macro precision, recall, and F1 score compared to KoBERT (Base, wCE), demonstrating that acquiring domain-specific knowledge is particularly beneficial for complex multi-class tasks. In contrast, DAPT had a negative impact on the macro F1 score when the standard cross-entropy loss was used.

3.5 Comparison with GPT-5-mini

GPT-5-mini achieved total accuracy comparable to the best KoBERT models. However, due to very low recall, it resulted in the lowest macro F1 score among all tested models. This suggests that for complex, domain-specific tasks, fine-tuning remains a more effective approach than using a general-purpose commercial LLM. We acknowledge that GPT-5-mini's performance might improve with expert prompt engineering.

4. Conclusion

4.1 Summary and Future Work

In this study, Domain-Adaptive Pre-Training (DAPT) was performed on KoBERT using a domain-specific corpus on radiation regulations. The model was then fine-tuned to classify sentences according to the relevant clauses of a specific act. Four configurations of the fine-tuned KoBERT—with and without DAPT and using either cross-entropy or weighted cross-entropy loss—was tested and their performance was compared to GPT-5-mini. Evaluation was focused on total accuracy, macro precision, macro recall, and macro F1 score. In terms of total accuracy, the best-performing models were KoBERT (Base, CE) and KoBERT (DAPT, CE), while KoBERT (DAPT, wCE) achieved the highest macro F1 score. Also, the positive effect of DAPT was observed only in models using the weighted cross-entropy loss function. GPT-5-mini achieved total accuracy similar to that of KoBERT (Base, wCE) and KoBERT (DAPT, wCE). However, in terms of macro F1 score, GPT-5-mini showed the lowest performance among all models, suggesting that fine-tuning remains an effective approach for handling complex text classification tasks in a specific domain. Therefore, fine-tuning a domain-specific language model could be a promising solution for text processing tasks within a specific domain, and fine-tuned model is expected to reduce the consumption of resources required for tasks involved in document processing related to *Act on Protective Action Guidelines Against Radiation in the Natural Environment*.

Future work will focus on accurately classifying sentences associated with two or more clauses, that is, multi-label classification and addressing class imbalance in the fine-tuning dataset.

ACKNOWLEDGEMENT

This research was supported by a grant from Korea Atomic Energy Research Institute (KAERI) (No. KAERI 522430-25).

REFERENCES

[1] Gururangan, Suchin, et al. "Don't stop pretraining: Adapt language models to domains and tasks." arXiv preprint arXiv:2004.10964 (2020).
[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019.
[3] GitHub. SKTBrain/KoBERT. https://github.com/SKTBrain/KoBERT
[4] Kudo, Taku, and John Richardson. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing." arXiv preprint arXiv:1808.06226 (2018).