ISSN 2093-0542

ICONI

KOREAN SOCIETY FOR INTERNET INFORMATION

The 17th International Conference on Internet

(ICONI 2025)

Dec. 14-17, 2025 Okinawa Convention Center, Okinawa, Japan

http://www.iconi.org

Proceedings of ICONI 2025

| Organized by |

Korean Society for Internet Information (KSII)

| Sponsored by |

OCVB

Contents

Improving KLUE Classification Performance based on Multi-Trial Evaluation

Jaeho Kim, Jinmo Yang, Jihoon Kong, Chaeyun Seo, and R. Young Chul Kim*
SE Lab, Hongik University
Sejong, South Korea
[e-mail: {jaehokim1005, yjmd2222}@g.hongik.ac.kr, go400s@gmail.com,
{chaeyun, *bob}@hongik.ac.kr]
*Corresponding author: R. Young Chul Kim

Abstract

KLUE is a benchmark for evaluating Korean Natural Language Understanding performance. Among these, Yonhap News Topic Classification (YNAT) is a representative text classification task. However, a significant hurdle in assessing model performance and ensuring fair comparisons is the absence of a standard evaluation dataset. To address this, we propose a robust evaluation framework based on our Multi-Trial evaluation mechanism and analyze the effects of syntactic data augmentation. First, we construct a statistically evaluated environment by repeating the process of splitting the training data at an 8:2 ratio with 20 different seeds. Using this framework, we compare the performance of six pre-trained language models (PLMs). Second, we augment news headlines into simple, compound, complex, and colloquial sentences, and then quantitatively analyze the impact of each sentence type on model performance improvement. Experimental results showed that the klue/roberta-large model, pre-trained in Korean, achieved the best performance, with an average Macro F1 score of 0.8703. In the experiments by sentence type, compound sentences showed the best performance with an average Macro F1 Score of 0.8636 when learning, and in the data augmentation experiments, the average Macro F1 scores for all types were highly similar at 0.863 and 0.8632. In conclusion, for the KLUE-YNAT task, we identified the most effective model architecture among the six tested and analyzed the impact of syntax-based text data augmentation techniques on the classification models.

Keywords: Korean Natural Language Topic Classification, Natural Language Processing

1. Introduction

Recently, Pre-trained Language Models (PLMs) based on the Transformer architecture have emerged as the preeminent and dominant paradigm for achieving state-of-the-art (SOTA) performance in the field of Natural Language Processing (NLP) [1, 2]. These PLMs are pre-trained on large-scale corpora to learn the statistical, grammatical, and semantic patterns of the language, positioning them in an optimal state for fine-tuning on specific downstream tasks [2]. In this trend, the Korean Language Understanding Evaluation (KLUE) benchmark emerged to advance Korean NLP technology and establish standardized evaluation criteria [3]. Yonhap News Topic Classification (YNAT), one of the several tasks in KLUE, is a core Topic Classification (TC) task that involves classifying

This research was supported by Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2025 (Project Name: Artificial Intelligence-based User Interactive Storytelling 3D Scene Authoring Technology Development, Project Number: RS-2023-00227917307820870502 01) and National Research Foundation (NRF), Korea, under project BK21 Four.

news headlines into seven topics, such as politics, economy, and society. However, the KLUE-YNAT task currently faces a critical limitation. Evaluation on the official test dataset was previously measurable only through leaderboard submissions, but the leaderboard's operation has now been suspended. This situation requires using the validation dataset as the test dataset and necessitates creating a new validation set by splitting the training dataset. This leads to evaluations that are dependent on data splits, thereby reducing reliability. Furthermore, given that not all PLMs exhibit identical performance on KLUE-YNAT TC, it is necessary to select the most suitable model architecture. Moreover, the YNAT dataset is confined to a limited data type, news headlines. This may limit its ability to demonstrate robust performance on novel types of headline sentences. This study conducts two experiments to address these problems. First, to overcome the problem of the absent official test set, we employ an evaluation methodology that ensures statistical validation by repeatedly splitting the training data at an 8:2 ratio with 20 different seeds. Through this approach, we compare the performance of six models in terms of average, maximum, and minimum scores, as well as standard deviation, and identify the optimal model architecture. Second, we analyze the impact of syntactic diversity by augmenting the dataset with structurally transformed headlines and evaluating their effect on model performance. Using this technique, we quantitatively demonstrate which type of sentence structures are learned most effectively by the six Transformer encoder-based classification models. The contributions of this paper can be summarized as follows:

A. Using a reliable and statistically evaluation method for the KLUE-YNAT task, we identified the model architecture that yields optimal performance among the six models tested.

B. Instead of using simple data augmentation, we analyzed the impact of syntax-based techniques by fine-tuning classification models on data augmented with specific sentence types.

2. Related Works

2.1 KLUE

The KLUE dataset was introduced to advance Korean NLP [3]. In addition to providing datasets for the benchmark, KLUE also released several versions of PLMs for future research. The KLUE benchmark evaluates performance in eight tasks. Among these, the KLUE-YNAT Topic Classification is a task that classifies Yonhap news headlines into seven topics. However, the operation of the KLUE leaderboard has been suspended, which makes final performance evaluation difficult.

2.2 Model Performance Evaluation

K-Fold Cross-Validation (K-Fold CV) is widely used to estimate the performance of a model. K-Fold CV involves dividing the entire dataset into K subsets, using K-1 subsets for training and the remaining one for validation, and repeating the process K times to calculate the average performance. This method provides high reliability for estimating the model's generalization performance. This study is inspired by K-Fold CV. It divides the data by setting a random seed and then fine-tunes the models.

2.3 Text Data Augmentation

The performance of a model is dependent on the quantity and quality of the training data. Text data augmentation is a key technique for addressing the problem of data scarcity. It can enhance model performance by generating new data that is semantically identical to the existing data but varied in expression. Prominent text data augmentation techniques include EDA [4] and Back-Translation [5]. While these two methods can achieve lexical and semantic diversity by replacing individual words or altering meanings, they have limitations in securing syntactic diversity. Yang et al. (2025) proposed a method for augmenting datasets by generating sentences that are semantically identical but have different syntactic structures [6]. Their method involved an LLM automatically transforming original sentences into four forms, simple, compound, complex, and colloquial via prompt engineering. This study applies the ideas proposed by Yang et al. to the fine-tuning stage of BERT-based models.

3. Methodology

3.1 Dataset

Two types of datasets are used in this study. These are the KLUE dataset, provided within the Transformers library, and augmented_klue.json, which stores the augmented dataset. The YNAT task provides a total of 45,678 training samples and 9,107 validation samples. The augmented_klue.json training data contains a total of 28,390 samples including title, simple, complex, compound, and colloquial. In this study, we use the provided validation data as the test data and split the original training data into new training and validation sets.

3.2 Model Training and Performance Measurement

We compared a total of six model types, google/bert-base-uncased, google/bert-base-multilingual, klue/bert-base, klue/roberta-large, klue/roberta-base, and klue/roberta-small. Since the KLUE benchmark does not provide an official test dataset, we split the 45,678 training samples 20 times at a random 8:2 ratio. After that, each of the six model types is trained on each of the 20 different data splits. We measure the F1 score for each of the 120 fine-tuned models. For each model type, we analyzed the average, maximum, minimum, and standard deviation from the results of the 20 distinct data splits. This is a method to enhance reliability and measure more generalized performance through multiple evaluations, rather than a single-split evaluation. The models trained after applying text data augmentation by sentence type are also trained on each of the 20 different data splits. In this process, we directly examine the average, maximum, minimum, and standard deviation. Through this, we identify which sentence type yields the best performance when used for augmentation.

Finally, the impact of data augmentation is quantified by training models on a 1:1 mixture of original and augmented data and comparing their performance to a baseline trained only on the original data.

4. Experiments

All experiments were conducted under identical hyperparameter settings, and the Macro F1 score was used as the performance metric. The hyperparameter values were set as follows:

- Learning rate: 2e-5
- Epochs: 3
- Batch size: 32
- Split ratio: 8:2
- Weight decay: 0.01

4.1 Performance Comparison of Pre-trained Models

The results of evaluating the six PLMs with 20 different seeds are shown in **Table 1**.

Table 1. Performance Comparison of Pre-trained Models

Model	Mean	Max	Min	Std Dev.
Bert-google-base	0.6519	0.6727	0.6368	0.0079
Bert-klue-base	0.8694	0.8721	0.8651	0.0018
Bert-base-multilingual	0.8238	0.8289	0.8201	0.0022
Roberta-klue-small	0.8636	0.8757	0.8606	0.0016
Roberta-klue-base	0.869	0.8717	0.8649	0.0017
Roberta-klue-large	0.8703	0.8742	0.8664	0.0018

The experimental results show that the klue/roberta-large model achieved the best performance, with a mean Macro F1 score of 0.8703 and a max Macro F1 score of 0.8742. Models specialized for the Korean language generally outperformed the baseline model, bert-base-multilingual, which suggests that the task is highly dependent on the linguistic characteristics of Korean.

4.2 Model Performance Comparison by Sentence Type

We selected the klue/Roberta-large model, which demonstrated the best performance in the “Performance Comparison of Pre-trained Models” from Section 4.1. The results from training the model separately on each of the four sentence types are presented in **Table 2**.

Table 2. Model Performance Comparison by Sentence Type

Data	Mean	Max	Min	Std Dev.
Original(Baseline)	0.8703	0.8742	0.8664	0.0018
Simple	0.8611	0.8662	0.8537	0.0033
Complex	0.8606	0.8652	0.8458	0.0041
Compound	0.8636	0.8678	0.8551	0.0032
Colloquial	0.8615	0.8661	0.8521	0.0030

4.3 Performance Comparison of Text Data Augmentation

Similarly, we selected klue/roberta-large for the same reason. **Table 3** presents an analysis of the effectiveness of the four text augmentation techniques. These results were generated from a model trained on a 1:1 mixture of the original dataset and data augmented by each type.

Table 3. Model Performance Comparison After Data Augmentation

Data	Mean	Max	Min	Std Dev.
Original(Baseline)	0.8703	0.8742	0.8664	0.0018
Original + Simple	0.863	0.8664	0.861	0.0015
Original + Complex	0.863	0.8671	0.8595	0.0022
Original + Compound	0.863	0.8668	0.8579	0.0024
Original + Colloquial	0.8632	0.8687	0.8564	0.0027

4.4 Reproducibility

As specified in the official PyTorch documentation, experimental results are influenced by several factors. The factors that affect reproducibility are as follows:

- Dropout behavior during seed setting
- The non-deterministic behavior of cuDNN algorithms
- The value of the CUBLAS_WORKSPACE_CONFIG environment variable
- The randomness of the DataLoader’s multi-process data loading algorithm

To ensure the reproducibility of the evaluation results, we recommend the following operational requirements:

- Python 3.11.9
- Torch 2.5.1_cu121
- Transformers 4.41.2
- Scikit-learn 1.5.1
- Numpy 1.26.4

In addition, experimental results may vary depending on the server environment. The server environment used in this study is shown in **Table 4**.

Table 4. Experimental Server Environment

OS	GPU	CPU	Memory	Storage
Ubuntu 22.04.5 Desktop * 1EA	2x NVIDIA H100NVL 96GB	2x Xeon Gold 6416H(18C 2.2G) – Intel Xeon 4th Sapphire Rapids Processor	DDR5 64GB RDIMM 4800MT/s (Total 256GB)	1x Gen5 U.2 NVMe 7.62TB

5. Discussion

5.1 Analysis of Measured Performance

First, the finding in Section 4.1 that klue/Roberta-large achieved the highest performance is consistent with expectations. This is interpreted as resulting from the model’s larger size and its pre-training on the large-scale Korean dataset, which allowed it to capture and classify the linguistic characteristics of Korean better than other models.

Second, the superior performance achieved by training on compound sentences, as shown in Section 4.2, warrants particular attention. News headlines are characterized by their need to be concise and informative. Consequently, the structure of compound sentences, which logically connects multiple clauses, may have been particularly effective.

Third, the decrease in performance after text data augmentation, as noted in Section 4.3, is the key finding of this study. This may have been caused by the following factors:

- The sentences generated via the LLM, while syntactically diverse, may have lost characteristics of the original news headlines, such as subtle nuances.
- Adding augmented data at a 1:1 ratio expanded the data distribution. This expansion may have made it more challenging for the model to capture the core features, resulting in a performance drop.

5.2 Limitations of the Study

This study did not delve into the qualitative details of the augmented data used. The study was also limited in that it applied a simple 1:1 data augmentation ratio rather than exploring various mixing ratios.

6. Conclusion

We partially addressed the absence of a test dataset in the KLUE-YNAT classification task by increasing the number of training dataset splits. Furthermore, we augmented the limited news headline data into simple, compound, complex, and colloquial forms, and then analyzed the impact of this augmented data on the models. Through evaluations using 20 different data split seeds, we experimentally demonstrated that the klue/roberta-large model outperformed the other five PLM types, achieving an average Macro F1 score of 0.8703 and a maximum of 0.8742. Additionally, we confirmed that when training separately by sentence type, training on data generated in the "Compound" form yielded relatively high performance, with an average Macro F1 score of 0.8636 and a maximum of 0.8678. Finally, we compared models trained on data augmented at a 1:1 ratio. The comparison showed that the model trained only on the original data performed best. It achieved an average Macro F1 score of 0.8703. We also confirmed that the other augmented cases yielded highly similar scores, average Macro F1 scores of 0.8630 and 0.8632.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol.1, pp.4171-4186, 2019.

[3] S. Park, et al., “KLUE: Korean Language Understanding Evaluation,” *arXiv preprint arXiv:2105.09680*. 2021.

[4] J. Wei and K. Zou, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks,” *arXiv preprint arXiv:1901.11196*, 2019.

[5] R. Sennrich, B. Haddow, and A. Birch, “Improving Neural Machine Translation Models with Monolingual Data,” *arXiv preprint arXiv:1511.06709*, 2015.

[6] J. Yang, C. Seo, K. Kim, J. Kim, and R. Y. C. Kim, “Topic Classification Training Model with Automatic Textual Data Transformation,” in *Proc. of International Conference on Green and Human Information Technology 2025*, pp.78-81, 2025.