Uncertainty-based Active Learning with Ensemble Technique for Enhancing the Performance of Natural Language Classification with Limited Data

Seong-Won Jeon
dept. of Applied Artificial Intelligence
Major in Bio Artificial Intelligence
Hanyang University
Ansan, Korea
seacom0601@hanyang.ac.kr

Dong-Ho Lee
dept. of Applied Artificial Intelligence
Major in Bio Artificial Intelligence
Hanyang University
Ansan, Korea
dhlee72@hanyang.ac.kr

Abstract—**Recently, advances in artificial intelligence have been rapidly driven by the development of large-scale language models, such as GPT-4. These models, trained on more extensive datasets, show remarkable performance across diverse natural language tasks. However, leveraging these models to create effective services can be resource-intensive. Particularly, in addition to the cost of refining and preprocessing data, getting a large amount of data and training them is very challenging. In this paper, we propose an uncertainty-based active learning approach with ensemble technique to enhance the performance of a natural language classification model using limited data. We achieve higher performance with less data regardless of data characteristics and the number of classes.**

Index Terms—**NLP, Active Learning, Uncertainty**

I. Introduction

Recently, there have been rapid advances in the field of artificial intelligence due to the development of large-scale language models such as GPT-4 [1]. These large-scale language models [2] [3], which are much larger in scale than previous language models, have been trained on a significantly large amount of data, resulting in improved performance across various natural language tasks. Various services are being developed by using these large-scale language models.

To develop services that provide good model performance, it is essential to fine-tune the model for specific tasks. However, using large-scale language models to create effective services is challenging. Training such models requires a substantial amount of data and computational resources, leading to concern about a huge cost. The process of refining and preprocessing an extensive amount of data consumes significant human resources and time. Working with domain-specific data or specialized fields, like healthcare, also involves costs and time for data validation.

Moreover, running large-scale language models typically requires enterprise-level hardware rather than standard personal equipment. For instance, GPT-3 [4] needs approximately 5,000 hours for training using 8 NVIDIA H100 designed for generative AI [5]. The cost of such equipment can be unaffordable for individual researchers. As a result, some researchers are focusing on fine-tuning small-scale language models to achieve better performance on specific tasks compared to large-scale language models. However, even in this process, having an ample amount of high-quality data remains crucial for achieving good performance. Therefore, reducing the cost of collecting and preprocessing the data emerges as a significant factor.

In this paper, we propose an uncertainty-based active learning approach with ensemble technique, Uncertainty Ensemble (UE), to achieve high performance with relatively less data by prioritizing the data that can have the most significant impact on improving the model's performance. Active learning based on uncertainty aims to strategically select a subset of data with the highest uncertainty, thus optimizing the cost of the learning process. By incorporating ensemble techniques with active learning, our approach leverages the strengths of variously defined uncertainty to address the challenges posed by scarce data. We use various natural language classification tasks as experimental scenarios. We also generate augmented datasets for each scenario, generated using rule-based augmentation techniques to minimize data validation costs. By using UE approach, we achieve higher performance with a smaller amount of data. Furthermore, regardless of data characteristics and the number of classes, there are significant performance improvements.

This work was partly supported by the Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)), the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No.NRF-2022R1F1A1073208) and Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0008691, HRD Program for Industrial Innovation).

This research (paper) used datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All data information can be accessed through 'AI-Hub (www.aihub.or.kr)'.

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

II. Related Works

A. Uncertainty-based Active Learning approach

Language models generally improve in performance as they are trained on more data. Therefore, many studies are being conducted to secure sufficient data with minimal cost.

However, it's not possible to infinitely add more data. While there is a limit to how much performance can be improved by adding more training data, the training cost can increase without bounds. To address this, many researchers study an active learning approach. Active learning approaches aim to reduce training costs while also minimizing data labeling expenses. It involves selecting data that would be most beneficial to the model's learning process, prioritizing data with high utility, and labeling them for model training. This approach involves selectively choosing the most impactful data to train the model. There are various approaches to select data that model needs to study first.

[6] is a publication that summarizes active learning approaches researched up to the time of its publication. It provides definitions of active learning approaches and presents various query strategies. Among them, a commonly used approach considers uncertainty, as higher uncertainty often leads to a greater learning effect. Defining uncertainty in different ways results in three distinct techniques:

First, Least Confidence [6] [7] defines higher uncertainty for data with lower top-1 prediction probabilities. This is because when the model predicts a label with low confidence, it indicates uncertainty about that label, thus we choose such data for priority in model's learning process.

Second, Margin Sampling [6] [8] defines higher uncertainty for data with smaller differences between the top-1 and top-2 prediction probabilities. This suggests that when the model struggles to distinguish between top-1 and top-2 and is uncertain about the correct label, the data is uncertain.

Third, Maximum Entropy [6] defines higher uncertainty based on the entropy of each data. Since entropy is used to measure the amount of information, data with higher entropy implies that the model has less knowledge about it. Thus, these data points are prioritized for model's learning process as they hold valuable information that the model needs to learn first.

However, all three approaches are vulnerable to outliers, leading to the exploration of alternative active learning approaches based on factors other than uncertainty [6] [9] [10].

First, Query by Committee [6] [9] uses ensemble approach to collect predictions from different models. Data with the most disagreement among the models is selected for priority in model's learning process. While it outperforms uncertainty-based approaches, the downside is the higher computational cost and workload of training multiple models. Moreover, limitations arise in models like BERT, where meaningful committee votes might be challenging to obtain.

Second, Core-set [10] prioritizes diverse data points by calculating distances between data points based on their convolutional features, representative data points are selected as a core-set to be trained first. The Core-set outperforms other active learning methods in image classification problems. However, there is a limitation in applying this approach to natural language classification problems.

In this paper, we propose an active learning approach to reduce the cost of learning and data processing. We ensemble uncertainty-based active learning approaches that minimize the increase in training costs. We define data prioritization according to the highest consensus of opinions on these uncertainty-based active learning approaches which implies a need for model's learning process. Additionally, we use this active learning approach on augmented data generated through rule-based data augmentation to prioritize and select sampled data. Finally, we estimate the learning cost changes by the time of validation of the sampled data and training with it.

B. Rule-based Data Augmentation technique

A model requires a substantial amount of data for good performance, which demands significant costs and time. Resources are also needed to refine and preprocess the data to make it suitable for training. In specialized fields like healthcare, there's an additional expenditure of time and effort for inspecting the final dataset. Consequently, rule-based augmentation techniques that minimize refinement and inspection costs are often favored over generative model-based augmentation techniques that need detailed inspection by many experts. [11] introduces rule-based data augmentation techniques for symptom expression data, considering specific linguistic characteristics of the Korean language. Therefore, in this paper, we generated augmented data using this rule-based augmentation technique to minimize data inspection costs.

III. Datasets

We use three Korean text datasets for this paper. These include two datasets from the medical domain, considered as a specialized field, and one dataset containing everyday conversations, considered as a more general field than the medical domain. These diverse datasets with varying characteristics are used in our experiments.

Table I shows the sizes of the base versions of the three datasets as well as the augmented data generated using rule-based augmentation techniques.

TABLE I
Sizes of Each Dataset

	sym_base	sym_aug	loc_base	loc_aug	conv_half	conv_aug
Train	21,860	382,167	18,786	328,757	43,631	390,709
Test	4,184	71,732	3,603	61,733	10,962	98,587

* 'sym' stands for 'symptom', 'loc' stands for 'symptom location', and 'conv' stands for 'conversation'.

A. Korean Symptom Expression Dataset

1) Symptom Classification

[11] constructed a patient symptom expression dataset for a total of 305 symptoms by excluding symptoms related to rare diseases, mental disorders, and pediatric conditions from the 1,005 symptoms provided by Asan Medical Center.

2) Symptom Location Classification

The symptom expression dataset presented in [11] had 305 classes, resulting in suboptimal classification performance. To address this, we create a scenario to train with fewer classes on the same data. Asan Medical Center provides information about the body parts where most symptoms occur. Thus, we make a mapping dictionary that replaced symptoms with corresponding symptom locations. Additionally, we make a subset of symptom-expression pairs containing symptom location information and conduct a process to replace symptoms with their corresponding symptom location. This led to the construction of a symptom expression dataset with a total of 11 categories of symptom location.

3) Rule-based Data Augmentation

We use the rule-based data augmentation technique proposed in [11] to make an augmented symptom expression dataset. Therefore, the augmented dataset is generated by using the repetition factor of 5.

B. Korean Everyday Conversation Dataset

1) Conversation Topic Classification

AIHub has released a Korean everyday conversation dataset [12] covering 20 conversation topics. To facilitate manageable training time, we randomly sample half of the total data and use it for training.

2) Rule-based Data Augmentation

We use the rule-based data augmentation technique proposed in [11]. Since the data length exceeded that of the symptom expression dataset, the rate of change was set to 0.1 and the repetition factor was set to 1.

IV. Methods

Active learning approach is one of the various strategies used to improve the model performance with lower resource. Active learning aims to reduce training costs while minimizing data processing expenses. Therefore, this approach focuses on prioritizing and sampling data that can offer the most significant performance improvement of the model.

A. Random Sampling (RS)

Random Sampling is an active learning approach which randomly extracts the sample dataset. As this approach doesn't have a specific sampling standard, it is often used as a baseline for comparing the performance of other active learning approaches. Thus, in each sampling iteration, 10,000 data are randomly extracted from the augmented dataset.

B. Uncertainty-based Sampling

Uncertainty-based Sampling is an active learning approach which uses uncertainty as a sampling standard. Thus, it can reduce the costs of training and data processing effectively.

1) Least Confidence (LC)

Least Confidence defines higher uncertainty for data with lower top-1 prediction probabilities by the model. Therefore, we predict the augmented dataset with a model trained on the base datasets and extract the top 10,000 data with the lowest top-1 prediction probabilities in each sampling iteration.

LC is a basic approach of uncertainty-based active learning approaches. Therefore, we use the LC's performance in experiments to compare with the performance of UE.

2) Margin Sampling (MS)

Margin Sampling considers higher uncertainty for data where the difference between the top-1 and top-2 prediction probabilities is smaller. We predict the augmented dataset with the same model we mentioned in LC and compute the difference between the top-1 and top-2 prediction probabilities. The top 10,000 data with the smallest differences are selected as the samples in each sampling iteration.

3) Maximum Entropy (ME)

Maximum Entropy defines higher uncertainty for data with higher entropy values. We predict the augmented dataset with the same model we mentioned in LC. The entropy for each data is calculated based on its prediction probabilities, and the top 10,000 data with the highest entropy values are chosen as the samples in each sampling iteration.

C. Uncertainty Ensemble (UE)

We propose a new active learning approach that combines uncertainty-based active learning approaches with ensemble technique to minimize the increase in training costs. The data to be prioritized for training are determined by the highest consensus of opinions from LC, MS, and ME. We extract 10,000 data points from each approach, remove duplicated data from the resulting 30,000 data, sort them based on the scores, summation of priority, and finally select the top 10,000 data in each sampling iteration.

Algorithm 1 explains UE approach in pseudocode. With a sample quantity($size$), set to 10,000, subsets of samples are drawn from uncertainty-based LC, MS, and ME. These three subsets are combined to create a $sample_{pool}$. The unique data values that appear in the $sample_{pool}$ are sorted in ascending order based on their occurrence frequency, and the top 10,000 data are selected to be used as UE samples.

Algorithm 1 Uncertainty Ensemble

```
Require: sample quantity $size$
    Base Dataset $D_{base}$, Augmented Dataset $D_{aug}$
    Least Confidence Function $f_{lc}$
    Margin Sampling Function $f_{ms}$
    Maximum Entropy Function $f_{me}$
    // All function returns sampled data which is extracted with
    each active learning approach.

procedure UNCERTAINTY ENSEMBLE($D_{base}, D_{aug}, size$)
    $D_{pool}$ = $D_{aug}$ - $D_{base}$
    $sample_{lc}$ = $f_{lc}(D_{pool}, size)$
    $sample_{ms}$ = $f_{ms}(D_{pool}, size)$
    $sample_{me}$ = $f_{me}(D_{pool}, size)$

    $sample_{pool}$ = $sample_{lc}$ + $sample_{ms}$ + $sample_{me}$

    $sample_{score}$ = $sample_{pool}$.value_counts()
    $sample_{sorted}$ = $sample_{score}$.sort_values(ascending=False)

    $sample_{data}$ = $sample_{sorted}$[:$size$]
    $D_{new\_train}$ = $D_{base}$ + $sample_{data}$

    return $D_{new\_train}$
```

V. Experiments

To analyze the impact of the active learning approaches mentioned in Section IV on the performance and training cost of natural language classification models, the following experiments were conducted by using KoBERT [13], a widely-used model for Korean language classification tasks.

A. Experimental Scenarios

1) Base Dataset

For each dataset, experiments were conducted by using the base datasets. The baseline performances are defined as the result of this experimental scenario.

2) Augmented Dataset

For each dataset, experiments were conducted by using the augmented datasets. The goal performance for active learning approaches is served as the result of this experimental scenario.

3) Active Learning approach

For each dataset, we define the data pool which has the data only from the augmented dataset, not from the base dataset. We extract 10,000 data from data pools using the RS, LC, and UE methods defined in Section IV. These samples are combined with the base dataset to form a new training dataset. Subsequently, we train the model used in the Base Dataset scenario with the new training dataset. The sampling and training processes are repeated until we achieve the goal performance of active learning approaches.

B. Results

We conducted 3 natural language classification experimental scenarios for each dataset. All experiments measured the model's performance based on the F1 score and the training cost based on the average training time per epoch and the total training time. All results are the average score or time of more than 5 repetitions of the same experiment. In the case of active learning approaches, the performance achieved with the augmented dataset was defined as the goal performance, and active learning sampling and training were repeated until the goal performance was achieved. For active learning, the training time included both the sampling and training processes, and this was defined as Active Learning time (AL time). Furthermore, instead of model epochs, 'sampling once and training until model converges' was defined as one active learning process, and the number of process iterations was defined as Active Learning count (AL count).

In all experiments, most of the total AL time was highest for RS, followed by LC and UE. RS takes the longest time because randomly extracted sample confuses the model. As LC has the outlier problem, the model was confused by the outlier data. But UE alleviate the outlier problem with ensemble technique, thus, UE's AL time was shortest in most experiments. Additionally, all three experiments achieved performance improvements with less than half of the augmented dataset, and the additional training cost was also less than half of the augmented dataset. These results indicate that active learning approaches can achieve good performance while reducing data processing and training costs. But RS and LC have many non-improvement results during the repetitions of the experiment, because of the random and outlier problem. However, based on these results, UE allows us to consider three uncertainties to select proper data for model's performance improvement, with less impact of outlier problem in uncertainty-based on active learning approaches. Therefore, UE is better than RS and LC for achieving both performance improvement and resource savings.

1) Symptom Classification

In both the Base Dataset and Augmented Dataset scenarios, symptom classification experiments showed F1 scores of 0.75 and 0.82, respectively, resulting in an improvement of approximately 0.07 with the augmented dataset. When using active learning approaches, RS, LC, and UE all achieved improved results of 0.78, 0.79, and 0.81, respectively, compared to the baseline performance. However, RS and LC did not achieve the goal performance of 0.82 and converged. RS and LC did not show significant improvement after the first active learning process, leading to a marginal improvement and convergence. UE demonstrated a stable performance improvement before convergence but required more AL time than RS.

TABLE II
RESULTS OF SYMPTOM CLASSIFICATION

	Base Dataset	Augmented Dataset	Random Sampling	Least Confidence	Uncertainty Ensemble
F1 score	0.75	0.82	0.78	0.79	0.81
Avg. training time/epoch	24M:22S	2H:37M:55S	-	-	-
Total training time	2D:17H:2M:20S	2D:20H:26M:8S	-	-	-
Avg. AL time/AL count	-	-	3H:49M:19S	2H:56M:31S	2H:19M:28S
Total AL time	-	-	7H:38M:38S	5H:53M:2S	9H:17M:52S
Total added data	-	-	20,000	20,000	40,000

TABLE III
RESULTS OF SYMPTOM LOCATION CLASSIFICATION

	Base Dataset	Augmented Dataset	Random Sampling	Least Confidence	Uncertainty Ensemble
F1 score	0.75	0.93	0.78	0.81	0.93
Avg. training time/epoch	20M:50S	3H:10M:53S	-	-	-
Total training time	9H:43M:21S	1D:14H:10M:37S	-	-	-
Avg. AL time/AL count	-	-	2H:38M:48S	2H:10M:20S	1H:38M:7S
Total AL time	-	-	7H:56M:24S	8H:41M:20S	6H:32M:28S
Total added data	-	-	30,000	40,000	40,000

TABLE IV
RESULTS OF CONVERSATION TOPIC CLASSIFICATION

	Base Dataset	Augmented Dataset	Random Sampling	Least Confidence	Uncertainty Ensemble
F1 score	0.41	0.48	0.44	0.44	0.48
Avg. training time/epoch	53M:36S	13H:18M:55S	-	-	-
Total training time	1D:1H:54M:48S	6D:15H:47M:3S	-	-	-
Avg. AL time/AL count	-	-	2H:32M:6S	2H:56M:56S	1H:41M:42S
Total AL time	-	-	10H:8M:24S	8H:51M:48S	8H:38M:30S
Total added data	-	-	40,000	30,000	50,000

2) Symptom Location Classification

For both the Base Dataset and Augmented Dataset scenarios in symptom location classification experiments, F1 scores of 0.75 and 0.93 were achieved, respectively, resulting in an improvement of approximately 0.18 with the augmented dataset. When using active learning approaches, RS, LC, and UE all achieved improved results of 0.78, 0.81, and 0.93, respectively, compared to the baseline performance. However, RS and LC did not achieve the target performance of 0.93 and converged. RS showed limited improvement after the first active learning process leading to marginal improvement and convergence. LC demonstrated stable performance improvement before convergence, requiring more AL time than RS. However, UE not only achieved the biggest improvements but required the shortest AL time among the three active learning approaches.

3) Conversation Topic Classification

For both the Base Dataset and Augmented Dataset scenarios in conversation topic classification experiments, F1 scores of 0.42 and 0.48 were achieved, respectively, resulting in an improvement of approximately 0.06 with the augmented dataset. When using active learning approaches, RS, LC, and UE all achieved improved results of 0.44, 0.44, and 0.48, respectively, compared to the baseline performance. However, RS and LC did not achieve the goal performance of 0.48 and converged. RS and LC showed limited improvement after the first active learning process, leading to convergence. UE demonstrated stable performance improvement before convergence, requiring less AL time than others.

VI. Conclusion

This paper proposes an active learning approach to address the increased data processing and training costs in natural language classification models. Active learning approaches extract the subset of data that has the most significant impact on model performance improvement from data pool. A new approach, Uncertainty Ensemble, applies an ensemble technique to uncertainty-based active learning approaches and aims to minimize training costs with less impact on the outlier problem in uncertainty-based active learning approaches. Experiments are conducted on various natural language classification tasks, including both specialized and general fields of Korean text datasets, as well as cases with varying numbers of classes. The results demonstrate that the UE approach consistently achieves significant performance improvements.

Therefore, even in scenarios with limited resources or data, and cases with a constrained number of classes, meaningful performance enhancements of the model can be achieved with minimal cost. However, while the ensemble technique helps alleviate the outlier problem, inefficiency due to outliers still exists. For future studies, we aim to explore the UE approach while considering data density to address this issue.

References

[1] R. OpenAI, "Gpt-4 technical report," *arXiv*, pp. 2303–08 774, 2023.

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[5] D. Khudia and V. Chiley, "Benchmarking large language models on nvidia h100 gpus with coreweave," (accessed Apr. 27, 2023). [Online]. Available: https://www.mosaicml.com/blog/coreweave-nvidia-h100-part-1

[6] B. Settles, "Active learning literature survey," 2009.

[7] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," in *Acm Sigir Forum*, vol. 29, no. 2. ACM New York, NY, USA, 1995, pp. 13–19.

[8] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *International symposium on intelligent data analysis*. Springer, 2001, pp. 309–318.

[9] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 287–294.

[10] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.

[11] S. W. Jeon, D. J. Lee, and D. H. Lee, "Construction of korean symptom articulation data using rule-based data augmentation technique," pp. 360–362, 2023.

[12] AIHub, "Textual everyday conversation data by topic."

[13] SKTBrain, "Kobert:korean bert pre-trained cased." (accessed 2019). [Online]. Available: https://github.com/SKTBrain/KoBERT