IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 July 2024, accepted 6 September 2024, date of publication 11 September 2024, date of current version 23 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3457850

RESEARCH ARTICLE

DSG-KD: Knowledge Distillation From Domain-Specific to General Language Models

SANGYEON CHO[1], (Graduate Student Member, IEEE), JANGYEONG JEON[1], (Member, IEEE), DONGJOON LEE[1], CHANGHEE LEE[2], (Member, IEEE), AND JUNYEONG KIM[1], (Member, IEEE)

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea
[2]Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea

Corresponding author: Junyeong Kim (junyeongkim@cau.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by Korea Government (MSIT) (Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics) under Grant 2022-0-00184; and in part by IITP Grant funded by Korea Government (MSIT) (Artificial Intelligence Graduate School Program, Chung-Ang University) under Grant 2021-0-01341.

ABSTRACT The use of pre-trained language models fine-tuned to address specific downstream tasks is a common approach in natural language processing (NLP). However, acquiring domain-specific knowledge via fine-tuning is challenging. Traditional methods involve pretraining language models using vast amounts of domain-specific data before fine-tuning for particular tasks. This study investigates emergency/non-emergency classification tasks based on electronic medical record (EMR) data obtained from pediatric emergency departments (PEDs) in Korea. Our findings reveal that existing domain-specific pre-trained language models underperform compared to general language models in handling N-lingual free-text data characteristics of non-English-speaking regions. To address these limitations, we propose a domain knowledge transfer methodology that leverages knowledge distillation to infuse general language models with domain-specific knowledge via fine-tuning. This study demonstrates the effective transfer of specialized knowledge between models by defining a general language model as the student model and a domain-specific pre-trained model as the teacher model. In particular, we address the complexities of EMR data obtained from PEDs in non-English-speaking regions, such as Korea, and demonstrate that the proposed method enhances classification performance in such contexts. The proposed methodology not only outperforms baseline models on Korean PED EMR data, but also promises broader applicability in various professional and technical domains. In future works, we intend to extend this methodology to include diverse non-English-speaking regions and address additional downstream tasks, with the aim of developing advanced model architectures using state-of-the-art KD techniques. The code is available in https://github.com/JoSangYeon/DSG-KD.

INDEX TERMS Bilingual medical data analysis, emergency room electronic health records, code switching, knowledge distillation, multilingual language models, natural language processing.

I. INTRODUCTION

With the proactive utilization of Electronic Medical Records (EMR) by healthcare institutions worldwide, a vast amount of medical information is being stored as data. [1], [2], [3] In particular, EMRs are documented in Pediatric Emergency Departments (PEDs) to capture patients' conditions at the time of visit, test results, and additional details in the form of free-text. These free-text entries are crucial components of EMRs and include essential clinical notes, such as the patient's gender, age, and vital signs. Given the importance of initial responses in PEDs, distinguishing between emergency

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

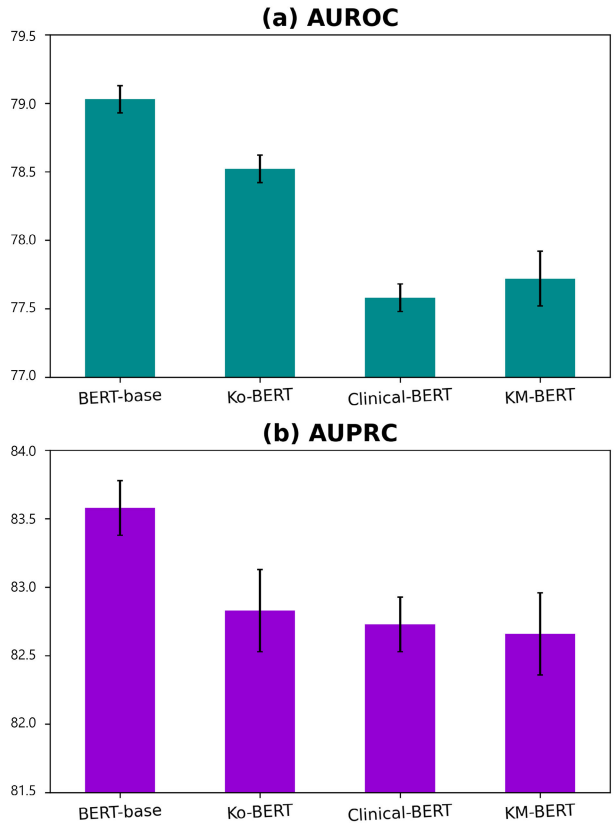


FIGURE 1. Generalized LM performs better. Performance of each pre-trained LM on an emergency/non-emergency classification task using EMR data from Korean PEDs in terms of (a) AUROC and (b) AUPRC.

and non-emergency patients based on EMRs recorded at the time of admission is critical [4], [5], [6]. However, owing to the nature of PED environments, clinical notes are often recorded in an unstructured and inconsistent manner, complicating the classification of emergency and non-emergency patients based on them [7], [8], [9]. This reduces the utility of the data in supporting decision-making by healthcare professionals, such as physicians [10], [11].

In particular, EMR data obtained from PEDs in Korea, a non-English-speaking country, are written in Korean. Non-English-speaking countries present an additional layer of complexity in classification tasks owing to the use of both English and local languages [12], [13]. In such cases, medical jargon and test results are often written in English or using English abbreviations, while the patient's condition and chief complaints are predominantly documented in the local language (Korean) [14], [15]. Thus, in non-English speaking countries, the data often come in the form of N-lingual free-text, with critical medical terms presented in English and other parts present in the local language. This form of data presents significant challenges in downstream tasks and during the extraction of meaningful words.

To overcome these problems, language models have been introduced based on transformers pre-trained on specific domain data to equip them with general domain knowledge [16], [17]. In particular, KM-BERT [18], pre-trained on Korean medical data, has been proposed to perform downstream tasks in the medical domain based on fine-tuning. However, KM-BERT does not perform well in the context of data collected in non-English-speaking countries, such as Korea. As depicted in Figure 1, in the task of classifying emergency and non-emergency cases using EMR data obtained from Korean PEDs, general language models, such as Ko-BERT and BERT-base, outperformed domain-specific pre-trained models, such as KM-BERT and Clinical-BERT, in terms of Area Under the Receiving Operating Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC) [19], [20]. This indicates that language models pre-trained with medical domain data are vulnerable to complications caused by N-lingual free-text characteristics of EMR data. The current study was envisioned to address this issue.

In this study, we address this problem by extracting domain knowledge for training using Knowledge Distillation (KD) [21]. We define an LM pre-trained with medical domain data as the teacher model and a general LM as the student model. Then, we extract medical knowledge from the former model and transfer it to the general latter model. We identify words containing medical knowledge in the input text and design the training to enable the student model to learn the teacher model's hidden states and attention matrices for these words. Our experiments demonstrate that the proposed method achieves effective knowledge transfer between LMs, yielding the best performance in classifying emergency and non-emergency cases based on PED EMR data.

II. RELATED WORKS

A. NLP FOR CLINICAL NOTES (EMR)

EMR [22], [23] represents digitally formatted data that are systematically collected and electronically stored to document patients' health information. In recent years, EMR data obtained from PEDs have garnered significant attention as the effective processing of such data is likely to enhance decision-making in clinical environments [4], [24], [25]. Various related studies have focused on the utilization, analysis, and decision-support methods of EMR data obtained from PEDs from multiple perspectives [26], [27], [28].

Early research on EMR data processing primarily targeted English-based data [29], applying traditional NLP methods such as Bag of Words (BoW) [30], [31], [32] and Term Frequency-Inverse Document Frequency (TF-IDF) [33], [34]. However, owing to the significant increase in the amount of acquired medical data in recent years, pre-trained LMs based on transformers have emerged [35], [36], and active research has been conducted on fine-tuning these models for various medical datasets. Notably, LMs such as KM-BERT and Clinical-BERT have been pre-trained on extensive medical data, exhibiting great potential in various NLP tasks in the medical domain [16], [37].

TABLE 1. Data Analysis for words: For the analysis of data on a word-by-word basis, the ratio of Korean:English:Other is used to signify the language distribution. In this case, it is approximately 0.43:0.23:0.33. The proportion of medical terms in Korean is 5%, while that in English is 20%, which is very high.

	Total number of words	Number of Korean words(medical words)	Number of English words(medical words)	Number of other words
Train	4,599,471	1,989,252 (99,197)	1,071,523 (220,221)	1,538,696
Dev	1,152,777	499,117 (24,590)	269,161 (55,156)	384,499
Test	1,444,904	624,343 (31,372)	335,464 (68,684)	485,097
total	**7,197,152**	**3,112,712 (155,159)**	**1,676,148 (344,061)**	**2,408,292**

However, the aforementioned LMs exhibit certain vulnerabilities in the case of N-lingual, free-text-formatted EMR data obtained from PED environments of non-English-speaking countries (see Figure 1). Given the importance of initial responses in PEDs, such issues are critical [38], [39]. This study describes methodologies capable of effectively learning and processing such N-lingual and free-text formatted data.

B. N-LINGUAL FREE-TEXT DATA

The processing of N-lingual and free-text data presents various motivations and challenges for NLP and deep learning. Early research on NLP focused primarily on English-based datasets [29]. However, data obtained from diverse countries has motivated the study of N-lingual processing [40], [41], [42], [43], leading to extensive research. This has led to the development of robust models capable of understanding the nuances of mixed languages to enhance the efficiency and accuracy of analyzing texts containing multiple languages [44]. Moreover, with the increase in online content, including social media posts, emails, news articles, and comments, the amount of unstructured free-text data has also increased [42], [45]. Consequently, numerous studies on NLP have aimed to extract meaningful insights from unstructured data and transform them into structured information. In particular, LM development in NLP currently focuses on maximizing the understanding of free text [46].

In non-English-speaking countries, the collection of various data types has led to the emergence of datasets that exhibit both N-lingual and free-text characteristics [47], [48]. This also constitutes a major topic of study within NLP. Indeed, research has been conducted on EMR data collected at PEDs in non-English-speaking countries, in which both N-lingual and free-text characteristics are prevalent [49], [50]. Existing research has identified critical issues with the performance of traditional LMs and NLP methods on such data. Unlike previous studies, we aim to enhance our understanding of data with N-lingual and free-text characteristics using KD.

C. KNOWLEDGE DISTILLATION

In this study, domain knowledge is transferred from the teacher model to the student model using KD [21]. Traditionally, KD is used in the NLP field to transfer knowledge from a larger teacher model to a smaller student model, with a focus on model size and lightweighting [51], [52]. Previous research approached this by having the student model mimic the teacher model's predictions using relatively simple methods. Recently, KD between transformer-based LMs has evolved to mimic the representation vectors of hidden states and attention matrices. This study, however, does not focus on model size and lightweighting, but rather on "extracting the unique knowledge possessed by the teacher model and transferring it to the student model." Although we use conventional distillation methods [51], [52], we first define the knowledge that the student model needs to learn and subsequently extract it from the teacher model for distillation. Detailed explanations are provided in [Section III].

By examining previous studies that applied KD from a knowledge transfer perspective rather than a lightweighting perspective, active research has been conducted on transferring knowledge from image models to language models [53], [54]; however, little research has been conducted on knowledge transfer between language models [55]. Existing studies on knowledge transfer between language models have focused on the quality of overall knowledge transfer, rather than domain-specific knowledge transfer. Our approach differs from those proposed in previous studies in that it focuses on the exchange of domain knowledge between language models in different domains.

III. PROPOSED METHOD

In this section, we formalize the central problem considered in this study and introduce the proposed methodology.

A. TEXT PREPROCESSING & LABELING

EMR data are obtained from Korean PEDs and then analyzed and preprocessed before model training. Even though South Korea is a non-English-speaking country, English is often used alongside local languages. This is reflected in the data obtained, which are recorded in the form of bilingual free-text clinical notes. The preprocessing and analysis methods, described later, are summarized in Table 1.

First, the bilingual free-text clinical notes are preprocessed by performing language and symbol distinction processing and new character removal while retaining the meaning of

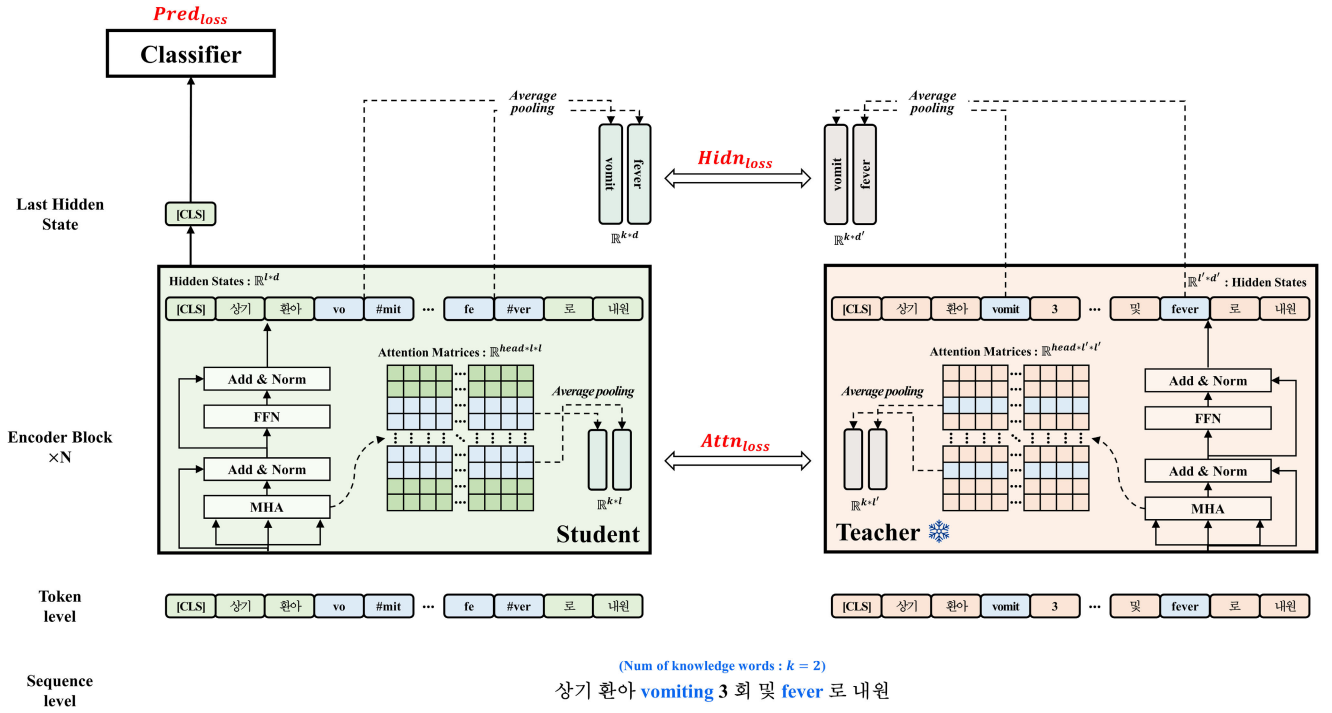


FIGURE 2. Model Architecture (Knowledge Transfer) - Visualization of the proposed methodology. The architecture consists of both student and teacher models based on transformers, which act as encoder blocks with multi-head attention (MHA) and feed-forward networks (FFN). The student model performs a prediction task ($\mathcal{L}_{\text{pred}}$) and receives appropriate domain knowledge from the teacher model by minimizing $\mathcal{L}_{\text{hidn}}$ and $\mathcal{L}_{\text{attn}}$. In the figure, we define "vomiting" and "fever" as domain knowledge words ($k = 2$) and perform distillation by receiving appropriate representations from the hidden states and attention matrices that arise from the teacher model. Overall, the goal of the proposed architecture is to transfer the teacher model's knowledge to the student, not only in terms of classification predictions but also throughout the model's internal representation, thereby enabling the student model to make decisions based on a deeper and more nuanced understanding of the input data.

the existing content as far as possible. Newline characters ($\backslash r, \backslash n$) generated during the process of recording and storing data in medical institutions are removed, and blanks are inserted between languages and symbols to separate them appropriately. This step is also performed for a method described below. Because we use a byte-pair encoding-based tokenizer, inserted blanks do not affect the input words severely [56]. In addition, because it is not possible to specify the features to be used as labels in the acquired data, active doctors are consulted to determine the label criteria using columns such as test status and medication history, as recorded in the data.

As per the preprocessed data, approximately 43% of the words are Korean, 23% are English, and 33% are symbols and other words. In addition, the number of English words in the data that are medical terms is evaluated by creating our own Korean medical dictionary by crawling the Korean Medical Search Engine and online medical dictionaries. According to Table 1, approximately 20% of the English words in the data are medical words, whereas only 5% of the Korean words are medical terms. This implies that almost all medical terms are more biased toward English than local languages and that medical domain knowledge is more embedded in English than in local languages.

B. METHOD

This section describes the domain knowledge transfer methodology proposed in this thesis in detail. The goal of this methodology is to transfer domain knowledge from a language model (teacher) pre-trained on a specific domain to a language model (student) pre-trained on general-purpose data. An overview of the framework is presented in Figure 2.

We define S as a generalized language model that categorizes urgent/non-urgent cases based on free-text clinical note data as input data. We train and validate the student model S. In addition, we define a domain-specific language model T that conveys domain knowledge and aids learning when S performs emergency or non-emergency classification tasks. The input free-text clinical notes are denoted by $X = \{x_i^S, x_i^T\}_{i=0}^N$. Here, x_i^S denotes the input of the S model, and x_i^T denotes the input of the T model. We also define $Y = \{y_i\}_{i=0}^N$, $Y \in \{0, 1\}$, and label it as emergency/non-emergency. Finally, we construct a dataset Data $= \{X, Y\}$ consisting of X and Y.

1) DEFINITION OF DOMAIN KNOWLEDGE

To transfer domain knowledge, we must define the domain knowledge present in the input x_i. This is defined as English words in the PED's EMR data (because, as discussed

previously, medical knowledge that the teacher needs to extract is almost always expressed in English, instead of the local language). To represent domain knowledge words (subwords) in the input sequence x_i, we represent the knowledge mask $\mathbf{M}$ as follows:

$$\begin{aligned} x_i &= \{t_1, t_2, \cdots, t_{l-1}, t_l\} \in \mathbb{R}^l \\ \mathbf{M}_i &= \{m_1, m_2, \cdots, m_{l-1}, m_l\} \in \{0, 1, \cdots, k\}^l \end{aligned} \quad (1)$$

where the knowledge mask $\mathbf{M}$ is $\mathbf{M} \in \mathbb{R}^l$, where each element m_l of $\mathbf{M}$ represents the index of the word for which the token at that position expresses a particular piece of domain knowledge. If $m_l = 0$, the token does not contain domain knowledge; if $m_l > 0$, the input x_i contains domain knowledge. In addition, k denotes the number of domain knowledge words. The appropriate domain knowledge token d_i is extracted from the input sequence x_i using $\mathbf{M}$.

$$d_i = \{x_i[j] | m_j > 0, \forall j \in \{0, \cdots, l\}\} \quad (2)$$

where d_i denotes the domain knowledge tokens corresponding to each token of the input x_i with the set relation $d_i \in x_i$.

Choosing the scope of domain knowledge definition is directly linked to model performance in the proposed methodology. In the data considered in this study, knowledge words are not defined as exclusively medical terms because the boundaries between medical and non-medical terms in the language are blurred. Thus, embedding the medical background knowledge of the teacher model in the English language itself is expected to benefit student learning. In particular, defining the appropriate domain knowledge based on the available data can be adopted as an inductive bias in the proposed domain knowledge transfer methodology, which is expected to yield better synergy.

In addition, the domain knowledge representation extracted from the teacher model must be defined. As we consider the language model of the transformer encoder series, we define the hidden states and attention matrices occurring in the encoder blocks of each layer as representations that imply domain knowledge. We distill the representation of English words observed in each input sequence x_i from the teacher to the student model.

2) PROBLEM FORMULATION

Let S and T contain P transformer layers. Knowledge transfer is achieved by performing KD on the encoder layers of the student and teacher models. S further solves the prediction task of classifying emergency/non-emergency cases, given X. Formally, the student receives domain knowledge from the teacher and solves the prediction task by minimizing the following objective function:

$$\begin{aligned} \mathcal{L}_{total} = &\sum_{x \in X} \mathcal{L}_{pred}(S(x^S), Y) \\ &+ \sum_{d \in X}^{M} \sum_{p=0}^{P} \mathcal{L}_{layer}(f_p^S(d^S), f_p^T(d^T), \lambda) \end{aligned} \quad (3)$$

where $\mathcal{L}_{pred}$ denotes the loss function that optimizes the classifier prediction of the student model performing the emergency/non-emergency classification task and $\mathcal{L}_{layer}$ denotes the loss function applied to a given layer of the model. Here, $f_p(\cdot)$ denotes the encoder block output in the p-th layer. The hyperparameter λ comprises $\lambda = \{\alpha, \beta\}$, where α and β are applied to $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$, respectively.

3) KNOWLEDGE DISTILLATION

In this section, we mathematically express the process by which a student model receives domain knowledge from a teacher model. To this end, two types of KD are performed: hidden-state and attention-matrix distillation [51], [52].

Hidden-state Loss is expressed as follows:

$$\mathcal{L}_{hidn} = \alpha \mathsf{MSE}(\hat{H}^S, \hat{H}^T) \quad (4)$$

Let H_S and H_T be the hidden states of the student and teacher encoder layers, respectively. Let H^S and H^T be $H^S \in \mathbb{R}^{l \times d}$ and $H^T \in \mathbb{R}^{l' \times d'}$, respectively. (Note that $d = d'$, $l = l'$).

$\hat{H}^S$ and $\hat{H}^T$ are expressed as follows:

$$\begin{aligned} \hat{H}^S &= \sum_{i=0}^{k} \text{Avg.Pool}(H^S_{m_i^S}, m^S) \in \mathbb{R}^{k \times d} \\ \hat{H}^T &= \sum_{i=0}^{k} \text{Avg.Pool}(H^T_{m_i^T}, m^T) \in \mathbb{R}^{k \times d'} \end{aligned} \quad (5)$$

Let m^S be a mask marking the domain knowledge words that the student model should receive from the teacher model. m^T is a mask marking the domain knowledge words of the teacher model. $m^S \in \mathbb{R}^l$ and $m^T \in \mathbb{R}^{l'}$.

Attention Matrices Loss is expressed as follows:

$$\mathcal{L}_{attn} = \beta \mathsf{MSE}(\hat{A}^S, \hat{A}^T) \quad (6)$$

Let A_S and A_T be the hidden states of the encoder layer for the student and teacher models, respectively. Let A^S and A^T be $A^S \in \mathbb{R}^{h \times l \times l}$ and $A^T \in \mathbb{R}^{h \times l \times l}$, respectively (but $l = l'$).

$\hat{A}^S$ and $\hat{A}^T$ are expressed as follows:

$$\begin{aligned} \hat{A}^S &= \sum_{i=0}^{k} \text{Avg.Pool}(A^S_{m_i^S}, m^S) \in \mathbb{R}^{k \times l} \\ \hat{A}^T &= \sum_{i=0}^{k} \text{Avg.Pool}(A^T_{m_i^T}, m^T) \in \mathbb{R}^{k \times l'} \end{aligned} \quad (7)$$

Finally, for each encoder layer, the loss is calculated using the following formula:

$$\mathcal{L}_{layer} = \begin{cases} \alpha \mathcal{L}_{hidn} & \text{n} = 0 \\ \alpha \mathcal{L}_{hidn} + \beta \mathcal{L}_{attn} & n \neq 0 \end{cases} \quad (8)$$

where $n = 0$ denotes the embedding layer and $\mathcal{L}_{attn}$ is not performed because no attention matrix is generated.

TABLE 2. Performance table: This table presents the performances of different models in terms of all metrics. The mean and standard deviation values are listed. The models are validated with respect to a variety of BoW-based topic models (Random Forest (RF), Logistic Regression (LR), XGBoost, and Gradient Boosting (GB)) and Transformer-based models (Ko-BERT, BERT-base, KM-BERT, Clinical-BERT, Bio-M-BERT, RoBERTa, and Bio-RoBERTa). The column entitled "The Proposed Method" lists the results obtained using the proposed methodology—each combination of student and teacher models is evaluated. The metrics used for evaluation are accuracy, area AUROC, AUPRC, recall, precision, and F1 Score, and the average values are provided in the Average column. In the column corresponding to each metric, the best performance is highlighted in boldface and the second-best performance is underlined.

		Accuracy	AUROC	AUPRC	Recall	Precision	F1 Score	Average
Ko-BERT		71.9±0.2	78.5±0.1	82.8±0.3	76.2±0.2	75.8±0.2	76.0±0.2	76.9±0.2
BERT-base		69.2±0.2	79.0±0.1	83.6±0.2	61.2±0.4	**81.7±0.2**	70.0±0.3	74.1±0.2
KM-BERT		71.0±0.1	77.7±0.2	82.7±0.3	77.9±0.2	74.0±0.3	75.9±0.2	76.5±0.2
Clinical-BERT		70.9±0.1	77.6±0.1	82.7±0.2	77.1±0.2	74.3±0.3	75.6±0.1	76.4±0.2
Bio-M-BERT		66.4±0.2	71.1±0.2	76.4±0.4	72.6±0.3	70.8±0.4	71.7±0.2	71.5±0.3
RoBerta		70.6±0.2	77.3±0.2	81.7±0.2	71.0±0.2	77.1±0.3	73.9±0.1	75.3±0.2
Bio-RoBerta		72.3±0.1	78.4±0.1	82.9±0.2	79.3±0.2	74.9±0.2	77.0±0.1	77.5±0.2
					ours			
Student	**Teacher**							
Ko-Bert								
	KM-BERT	72.6±0.4	79.4±0.4	84.1±0.4	**85.7±0.2**	72.7±0.5	**78.6±0.3**	**78.9±0.4**
	Clinical-BERT	72.6±0.2	79.4±0.2	83.8±0.2	74.6±0.3	77.8±0.2	76.2±0.2	77.4±0.2
	Bio-M-BERT	**73.4±0.2**	**80.4±0.2**	84.3±0.2	75.1±0.3	78.6±0.2	76.8±0.2	78.1±0.2
	RoBerta	73.1±0.2	80.1±0.2	84.6±0.2	77.7±0.3	76.7±0.2	77.2±0.2	78.3±0.2
	Bio-RoBerta	73.3±0.2	**80.4±0.2**	**85.0±0.2**	77.6±0.3	76.9±0.2	77.3±0.2	78.4±0.2
Bert								
	KM-BERT	70.9±0.2	77.1±0.1	82.3±0.2	79.6±0.3	73.2±0.3	76.2±0.2	76.5±0.2
	Clinical-BERT	70.8±0.2	76.7±0.2	81.6±0.3	74.4±0.3	75.4±0.2	74.9±0.2	75.6±0.2
	Bio-M-BERT	70.4±0.1	76.5±0.2	81.6±0.2	77.3±0.2	73.6±0.2	75.4±0.2	75.8±0.2
	RoBerta	70.8±0.2	76.9±0.2	82.0±0.2	84.6±0.2	71.1±0.2	77.2±0.1	77.1±0.2
	Bio-RoBerta	68.6±0.2	76.3±0.2	81.2±0.3	64.6±0.4	80.0±0.3	70.7±0.3	73.2±0.3

4) PREDICTION LOSS

The student model continues to perform and optimize the classification task.

$$\mathcal{L}_{pred} = \mathsf{CE}(S(x^S), Y) \tag{9}$$

where x^S denotes the input to the student model and $S(x^S)$ denotes the logits of the student model. $\mathsf{CE}(\cdot)$ denotes the cross entropy loss and Y is the label.

IV. EXPERIMENTS

A. DATASET

The proposed methodology is validated using EMR data obtained from Korean PEDs. The task considered in this study is to classify the data into emergencies and non-emergencies using binary classification.

B. IMPLEMENTATION DETAILS

In this section, we introduce the experimental environment and models on which the proposed methodology is applied. The A6000 GPU device is used, with a batch size of 32, 15 epochs of training with an early termination condition at 15 to choose the weight with the best valid performance. The learning rate is set to 1e-5 for the BERT body and 1e-2 for the classifier. For reproducibility, we set the random seed to 42.

The transformer-based language models Ko-BERT and BERT-base are used as baseline student models, and KM-BERT, Clinical-BERT (C-BERT), RoBerta, Bio-RoBerta (B-RoBerta), and Bio-Multilingual-BERT (Bio-M-BERT) are used as baseline teacher models. In addition, four BoW-based machine learning models (RF, LR, XGBoost, and GB) are trained and compared.

The transformer-based models have M=12, d=768, l=512, and h=12, where M denotes the number of encoder layers, d denotes the dimension of hidden state, l denotes the maximum sequence length, and h denotes the number of multi-headers. The classification performances of the transformer models trained and measured on the test set are compared with that of the proposed method with all possible combinations of student-teacher models while training on the test set. In the proposed methodology, knowledge transfer and classification are performed simultaneously.

C. RESULTS

Table 2 presents a comparison of all methods in different groups in terms of accuracy, AUROC, AUPRC, recall, precision, and F1 Score. In addition, the average values are listed in the Average column to facilitate the overall performance comparison. In general, transformer-based NLP models are observed to outperform traditional NLP processes proposed in previous studies. Transformer-based language models can adapt quickly to the language patterns present in data when fine-tuned using language patterns during pretraining.

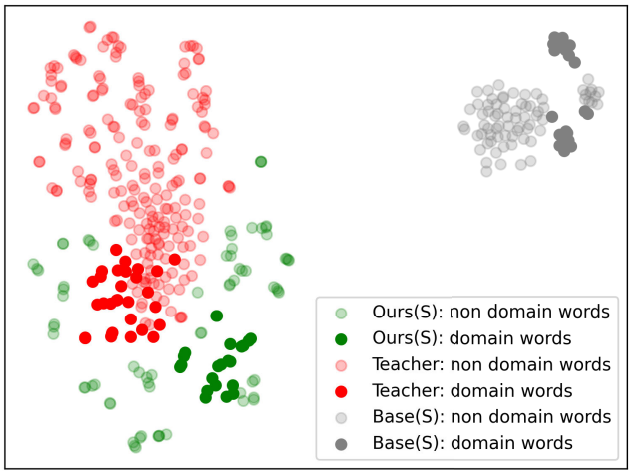


FIGURE 3. Qualitative analysis: Visual representation of the effectiveness of the proposed training method. S denotes a student model, which is taken to be Ko-BERT, and the teacher is taken to be KM-BERT. The gray dots correspond to an independently trained student model, the green dots correspond to the proposed training methodology, and the red dots correspond to the spatial coordinates of the teacher model's representation.

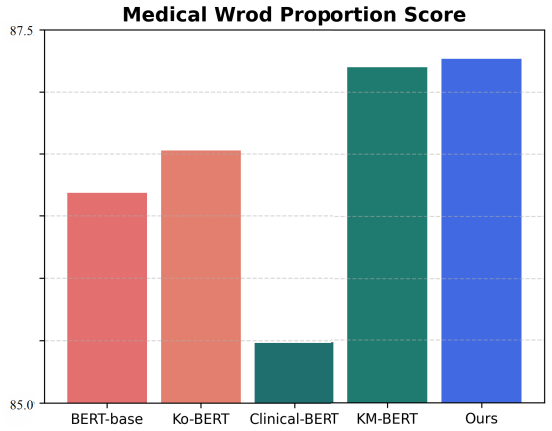


FIGURE 4. Case analysis: This figure shows calculated MWPS for correctly classified cases in the test set for each model. The proposed model (Ours) uses Ko-BERT as the student model and KM-BERT as the teacher model. By achieving the highest MWPS performance compared to all other models, our methodology demonstrates an effective understanding of medical words. Notably, the higher MWPS compared to the teacher model is particularly encouraging in terms of domain knowledge comprehension.

In particular, it is interesting to note from Table 2 that language models pre-trained with data from the medical field (KM-BERT, Clinical-BERT, etc.) perform worse than general language models (BERT-based, Ko-BERT) in terms of AUROC and AUPRC. This is because, as explained earlier, general language models perform favorably with respect to N-lingual and unstructured data while pretraining with a large amount of data, whereas language models pre-trained on the medical domain struggle on N-lingual or unstructured real data, even though they have learned medical domain information.

The bottom part of Table 2 presents the results obtained by training the existing language models using the proposed methodology. The student models are divided into Ko-BERT- and BERT-based cases and teacher models are taken to be pre-trained language models in the medical domain. The results are observed to be significantly better than those obtained by fine-tuning a simple language model. Moreover, the performance is even better than that of existing student and teacher models when trained alone. The aforementioned experimental results demonstrate the effectiveness of knowledge transfer between language models, leading to efficient interaction between knowledge of different language models.

1) QUALITATIVE ANALYSIS

To verify the relationship between domain words and teacher embeddings before and after DSG-KD visually, we present a scatter plot in Figure 3. Random samples are obtained from the dataset and visualized. In the figure, Ko-BERT is used as the student model and KM-BERT is used as the teacher model.

First, for each color, the darker shade represents the coordinates of the embedding vectors for words that contain domain knowledge, and the lighter shade represents the coordinates of the embedding vectors for words that do not. The gray points represent the coordinates of the embedding space when the student model is trained independently. The green points represent the coordinates of the embedding space of the student model trained using the proposed methodology. The red points represent the coordinates of the embedding space corresponding to the teacher model, which provides a good understanding of the medical domain. Visually, the embedding space of the student model trained using the proposed methodology is similar to that of the teacher model, which is rich in medical expressions. In contrast, when the student model is trained independently, it yields a relatively poor representation space and its visual representation is located far from that of the teacher model.

2) CASE ANALYSIS

Furthermore, to understand the impact of the proposed methodology on Domain words (in this study, medical words), we defined and measured the Medical Word Proportion Score (MWPS). The MWPS is defined as follows:

$$\text{MWPS} = \frac{\%\text{ of medical terms in English words}}{\%\text{ of English words in all words}} \tag{10}$$

This can be expressed mathematically as:

$$\text{MWPS} = \frac{\frac{m}{E}}{\frac{E}{A}} = \frac{m \cdot A}{E^2} \tag{11}$$

where m is the number of medical terms in English words, E is the total number of English words, and A is the total number of words.

We measured the MWPS for correctly classified cases in the test set for each model. This allowed us to evaluate the effectiveness of the proposed methodology. The comparison

of MWPS for each model is shown in Figure 4. From the figure, we can observe that our model achieves the highest MWPS, indicating that the DSG-KD methodology effectively learns Domain words, i.e., Domain Knowledge. Specifically, it demonstrates that medical knowledge can be transferred to a general language model and that our model shows a slightly higher understanding of medical terms than KM-BERT. In summary, this suggests that the proposed methodology is highly beneficial for applications in specific domains such as the medical field. Thus, it is proven that the DSG-KD methodology learns the medical knowledge embedded in medical terms and achieves superior classification performance.

D. ABLATION STUDY

The influence of $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ on performance is investigated by considering Equation 8. The influences of α and β are also studied. Two additional experiments are performed with Ko-BERT as the student model and KM-BERT as the teacher model. The performances are evaluated in terms of three metrics: AUROC, AUPRC, and F1 Score.

1) INFLUENCE OF $\mathcal{L}_{HIDN}$ & $\mathcal{L}_{ATTN}$ ON THE OBJECTIVE FUNCTION

Table 3 presents the effects of $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ on $\mathcal{L}_{total}$. The baseline model, which includes all loss components, exhibits an AUROC of 79.4±0.4, an AUPRC of 84.1±0.4, and an F1 Score of 78.6±0.3. This is used as the baseline for comparison in the ablation study, which is performed by removing $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ individually, or in combination. The results are as follows.

Case 1: Removing both $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$: When $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ are removed, all three performance metrics decrease, indicating that the two losses play significant roles in improving model performance. Case 2: Removing only $\mathcal{L}_{hidn}$: The absence of $\mathcal{L}_{hidn}$ also results in a decrease in performance, although to a lesser extent than in Case 1. This suggests that hidden losses contribute to model performance to a certain extent. Case 3: Removing $\mathcal{L}_{attn}$: The absence of $\mathcal{L}_{attn}$ has the largest effect on F1 Score, suggesting that this factor is particularly important for the accuracy and recall balance of the model. In contrast to Cases 2 and 3, removing both $\mathcal{L}_{hidn}$ and $\mathcal{L}_{attn}$ further degrades the performance, emphasizing the importance of synergistic functioning of these components to achieve high model performance.

This study demonstrates that each component plays an important role, and that prediction loss is the most important factor in maintaining the integrity of model performance. The cumulative degradation caused by removing multiple components suggests that the interactions among these components are complex and essential for high-quality model learning.

2) EFFECT OF α & β ON THE OBJECTIVE FUNCTION

Table 4 presents the impact of α and β on the performance, expressed in terms of performance. When $\alpha = 0.6$ and $\beta = 0.2$, AUROC = 79.4±0.4, AUPRC = 84.1±0.4, and F1 Score = 78.6±0.3 are observed. We systematically vary α and β to observe changes in the performance, and the results are presented below.

TABLE 3. Ablation study on objective function for $\mathcal{L}_{hidn}$, $\mathcal{L}_{attn}$.

$\mathcal{L}_{pred}$	$\mathcal{L}_{hidn}$	$\mathcal{L}_{attn}$	AUROC	AUPRC	F1 Score
✓	✓	✓	**79.4±0.4**	**84.1±0.4**	**78.6±0.3**
✓	-	-	78.5±0.1	82.8±0.3	76.0±0.2
✓	✓	-	79.2±0.3	83.9±0.3	78.6±0.2
✓	-	✓	78.1±0.3	83.1±0.4	67.8±0.3

Case 1: Impact of equal weights ($\alpha = \beta = 1.0$): Assigning equal weights to both hyperparameters results in a slight decrease in all metrics, suggesting that the model may not need to emphasize the factors controlled by α and β equally. Case 2: Increasing the emphasis on β ($\alpha = 1.0$, $\beta = 0.6$): Increasing β relative to α slightly improves the AUROC to 79.0±0.4, indicating that aspects of the model influenced by β may be more important for predictive ability. Case 3: Gradual adjustment of α and β: A pattern is observed when proportionally adjusting β from 1.0 to 0.1 while gradually decreasing α from 1.0 to 0.1, with the highest F1 Score achieved when $\alpha = 0.8$ and $\beta = 0.4$, and the best AUROC and AUPRC at $\alpha = 0.8$ and $\beta = 0.4$. This suggests an optimal range for the hyperparameters that balances the factors controlled by the model. Case 4: Significant reduction of α ($\alpha = 0.1$, $\beta = 0.1$): Significantly reducing both α and β to 0.1 results in the lowest AUROC, a slight increase in AUPRC, and a decrease in the F1 Score, which suggests that very low values of these hyperparameters can degrade the performance of the model, especially in terms of precision and recall balance.

In conclusion, the hyperparameters α and β play important roles in the performance of the Ko-BERT + KM-BERT model. For each data environment, an optimal parameter range yields the best performance. This emphasizes the importance of fine-tuning hyperparameters to satisfy specific performance goals during model training.

TABLE 4. Ablation study on objective function for α and β.

α	β	AUROC	AUPRC	F1 Score
1.0	**1.0**	78.8±0.4	83.4±0.4	78.2±0.3
1.0	**0.6**	79.0±0.4	83.5±0.4	75.1±0.4
0.9	**0.5**	78.5±0.4	83.1±0.4	78.3±0.3
0.8	**0.4**	79.0±0.4	83.6±0.4	78.4±0.3
0.7	**0.3**	79.1±0.4	83.6±0.3	77.9±0.4
0.6	**0.2**	**79.4±0.4**	**84.1±0.4**	**78.6±0.3**
0.1	**0.1**	77.8±0.3	81.7±0.5	76.9±0.3

V. CONCLUSION

This study addresses the need for effective domain knowledge transfer in pre-trained language models and

provides a promising approach for bridging the gap between general-purpose language understanding and domain-specific expertise. As the demand for domain expertise in NLP continues to increase, this study makes a timely and valuable contribution to improving the ability of language models to utilize specific and nuanced domain knowledge.

To this end, we introduce a novel methodology to extract domain knowledge from domain-specific pre-trained language models and transfer the knowledge to more generalized language models. We find that domain-specific pre-trained language models are vulnerable to bilingual data in non-English-speaking countries such as Korea, and demonstrate that by transferring domain knowledge to generalized language models, the proposed methodology is robust in N-lingual environments and exhibits effective classification performance in specific domains.

In this study, a highest average metric of 789±0.4 is achieved by transferring medical knowledge from the teacher model (KM-BERT) to the student model (Ko-BERT) using Korean emergency room EMR data. This demonstrates that the proposed framework can be used as a tool to assist doctors and healthcare workers in decision-making by helping them overcome N-lingual challenges inherent in EMR data recorded in non-English speaking countries. Beyond the medical field, the methodology can also be applied in various professional and technical fields, opening up potential avenues for its use as a decision aid in many forms.

In future works, we intend to extend and improve this methodology by utilizing EMR data obtained from diverse non-English-speaking countries, or by considering downstream tasks from various technical and professional fields in specific non-English-speaking countries. We also aim to develop an advanced model architecture by applying the latest KD techniques to deep learning methods.

REFERENCES

[1] N. Menachemi and T. H. Collum, "Benefits and drawbacks of electronic health record systems," *Risk Manage. Healthcare Policy*, vol. 4, pp. 47–55, May 2011.

[2] M. Hartswood, R. Procter, M. Rouncefield, and R. Slack, "Making a case in medical work: Implications for the electronic medical record," *Comput. Supported Cooperat. Work*, vol. 12, no. 3, pp. 241–266, Sep. 2003.

[3] F. Williams and S. Boren, "The role of the electronic medical record (EMR) in care delivery development in developing countries: A systematic review," *J. Innov. Health Informat.*, vol. 16, no. 2, pp. 139–145, Jul. 2008.

[4] W. S. Hong, A. D. Haimovich, and R. A. Taylor, "Predicting hospital admission at emergency department triage using machine learning," *PLoS ONE*, vol. 13, no. 7, Jul. 2018, Art. no. e0201016.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–10.

[6] V. Lambert, A. Matthews, R. MacDonell, and J. Fitzsimons, "Paediatric early warning systems for detecting and responding to clinical deterioration in children: A systematic review," *BMJ Open*, vol. 7, no. 3, Mar. 2017, Art. no. e014497.

[7] C. J. Harrison and C. J. Sidey-Gibbons, "Machine learning in medicine: A practical introduction to natural language processing," *BMC Med. Res. Methodol.*, vol. 21, no. 1, p. 158, Dec. 2021.

[8] Y. Zhang et al., "High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP)," *Nature Protocols*, vol. 14, no. 12, pp. 3426–3444, Dec. 2019.

[9] V. M. Castro, J. Minnier, S. N. Murphy, I. Kohane, S. E. Churchill, V. Gainer, T. Cai, A. G. Hoffnagle, Y. Dai, S. Block, S. R. Weill, M. Nadal-Vicens, A. R. Pollastri, J. N. Rosenquist, S. Goryachev, D. Ongur, P. Sklar, R. H. Perlis, and J. W. Smoller, "Validation of electronic health record phenotyping of bipolar disorder cases and controls," *Amer. J. Psychiatry*, vol. 172, no. 4, pp. 363–372, Apr. 2015.

[10] K. Adnan, R. Akbar, S. W. Khor, and A. B. A. Ali, "Role and challenges of unstructured big data in healthcare," in *Proc. Int. Conf. Data Manag., Anal. Innov. (ICDMAI)*, vol. 1, 2020, pp. 301–323.

[11] M. Tayefi, P. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen, "Challenges and opportunities beyond structured data in analysis of electronic health records," *WIREs Comput. Statist.*, vol. 13, no. 6, p. 1549, Nov. 2021.

[12] M. A. Gianfrancesco and N. D. Goldstein, "A narrative review on the validity of electronic health record-based research in epidemiology," *BMC Med. Res. Methodol.*, vol. 21, no. 1, pp. 1–10, Dec. 2021.

[13] E. Barbazza, S. Allin, M. Byrnes, A. D. Foebel, T. Khan, P. Sidhom, N. S. Klazinga, and D. S. Kringos, "The current and potential uses of electronic medical record (EMR) data for primary health care performance measurement in the Canadian context: A qualitative analysis," *BMC Health Services Res.*, vol. 21, no. 1, pp. 1–11, Dec. 2021.

[14] Y.-T. Park and D. Han, "Current status of electronic medical record systems in hospitals and clinics in Korea," *Healthcare Informat. Res.*, vol. 23, no. 3, pp. 189–198, 2017.

[15] A. Shinozaki, "Electronic medical records and machine learning in approaches to drug development," in *Artificial Intelligence in Oncology Drug Discovery and Development*. London, U.K.: IntechOpen, 2020.

[16] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ Digit. Med.*, vol. 4, no. 1, p. 86, May 2021.

[17] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*.

[18] Y. Kim, J.-H. Kim, J. M. Lee, M. J. Jang, Y. J. Yum, S. Kim, U. Shin, Y.-M. Kim, H. J. Joo, and S. Song, "A pre-trained BERT for Korean medical natural language processing," *Sci. Rep.*, vol. 12, no. 1, p. 13847, Aug. 2022.

[19] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *J. Biomed. Informat.*, vol. 38, no. 5, pp. 404–415, Oct. 2005.

[20] P. Flach and M. Kull, "Precision-recall-gain curves: PR analysis done right," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[22] S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural language processing of clinical notes on chronic diseases: Systematic review," *JMIR Med. Informat.*, vol. 7, no. 2, Apr. 2019, Art. no. e12239.

[23] O. Ben-Assuli, I. Shabtai, M. Leshno, and S. Hill, "EHR in emergency rooms: Exploring the effect of key information components on main complaints," *J. Med. Syst.*, vol. 38, no. 4, pp. 1–8, Apr. 2014.

[24] J. E. Anderson and D. C. Chang, "Using electronic health records for surgical quality improvement in the era of big data," *JAMA Surg.*, vol. 150, no. 1, pp. 24–29, Jan. 2015.

[25] A. Kirubarajan, A. Taher, S. Khan, and S. Masood, "Artificial intelligence in emergency medicine: A scoping review," *J. Amer. College Emergency Physicians Open*, vol. 1, no. 6, pp. 1691–1702, Dec. 2020.

[26] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Med. Informat. Decis. Making*, vol. 17, no. 1, pp. 1–13, Dec. 2017.

[27] I. Spasic and G. Nenadic, "Clinical text data in machine learning: Systematic review," *JMIR Med. Informat.*, vol. 8, no. 3, Mar. 2020, Art. no. e17984, doi: 10.2196/17984.

[28] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. M. U. Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," *Int. J. Inf. Technol.*, vol. 12, no. 3, pp. 731–739, Sep. 2020.

[29] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *Proc. Int. Conf. Text, Speech Dialogue (TSD)*. Pilsen, Czech Republic: Springer, 2007, pp. 196–205.

[30] G. Zuccon, A. S. Wagholikar, A. N. Nguyen, L. Butt, K. Chu, S. Martin, and J. Greenslade, "Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed CT ontology," *AMIA Summits Transl. Sci. Proc.*, vol. 2013, p. 300, Mar. 2013.
[31] R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, K. Jung, P. LePendu, and N. H. Shah, "Text mining for adverse drug events: The promise, challenges, and state of the art," *Drug Saf.*, vol. 37, no. 10, pp. 777–790, Oct. 2014.
[32] R. J. Delahanty, J. Alvarez, L. M. Flynn, R. L. Sherwin, and S. S. Jones, "Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis," *Ann. Emergency Med.*, vol. 73, no. 4, pp. 334–344, Apr. 2019.
[33] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Document.*, vol. 28, no. 1, pp. 11–21, 1972.
[34] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.
[35] L. Medrouk and A. Pappa, "Deep learning model for sentiment analysis in multi-lingual corpus," in *Proc. 24th Int. Conf. Neural Inf. Process. (ICONIP)*, Guangzhou, China. Springer, Nov. 2017, pp. 205–212.
[36] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu, "Deep learning in clinical natural language processing: A methodical review," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 3, pp. 457–470, Mar. 2020.
[37] X. Yu, W. Hu, S. Lu, X. Sun, and Z. Yuan, "BioBERT based named entity recognition in electronic medical record," in *Proc. 10th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Aug. 2019, pp. 49–52.
[38] S. Fuchs, M. Terry, K. Adelgais, M. Bokholdt, J. Brice, K. M. Brown, A. Cooper, M. E. Fallat, K. E. Remick, K. Widmeier, W. Simon, and M. Marx, "Definitions and assessment approaches for emergency medical services for children," *Pediatrics*, vol. 138, no. 6, Dec. 2016, Art. no. e20161073.
[39] H. Tohira, J. Finn, S. Ball, D. Brink, and P. Buzzacott, "Machine learning and natural language processing to identify falls in electronic patient care records from ambulance attendances," *Informat. Health Social Care*, vol. 47, no. 4, pp. 403–413, Oct. 2022.
[40] J.-E. Park, "Korean/English intrasentential code-switching: Matrix language assignment and linguistic constraints," Ph.D. dissertation, Dept. Educ., Univ. Illinois Urbana, Champaign, IL, USA, 1990.
[41] M. Nishimura, "Intrasentential codeswitching in Japanese and English," Ph.D. dissertation, Dept. Linguistics, Univ. Pennsylvania, Philadelphia, PA, USA, 1985.
[42] J. Ahn, C. L. Ferle, and D. Lee, "Language and advertising effectiveness: Code-switching in the Korean marketplace," *Int. J. Advertising*, vol. 36, no. 3, pp. 477–495, May 2017.
[43] D. Amazouz, M. Adda-Decker, and L. Lamel, "Addressing code-switching in French/Algerian Arabic speech," in *Proc. Interspeech*, Aug. 2017, pp. 62–66.
[44] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" 2019, *arXiv:1906.01502*.
[45] C. Baker, *Foundations of Bilingual Education and Bilingualism*. Bristol, U.K.: Multilingual Matters, 2011.
[46] J. Lee, "KcBERT: Korean comments bert," in *Proc. 32nd Annu. Conf. Hum. Cogn. Lang. Technol.*, 2020, pp. 437–440.
[47] M. Chen, W. Shi, B. Zhou, and D. Roth, "Cross-lingual entity alignment with incidental supervision," 2020, *arXiv:2005.00171*.
[48] H. Jin, C. Li, J. Zhang, L. Hou, J. Li, and P. Zhang, "XLORE2: Large-scale cross-lingual knowledge graph construction and application," *Data Intell.*, vol. 1, no. 1, pp. 77–98, Mar. 2019.
[49] Y. S. Bae, K. H. Kim, H. K. Kim, S. W. Choi, T. Ko, H. H. Seo, H.-Y. Lee, and H. Jeon, "Keyword extraction algorithm for classifying smoking status from unstructured bilingual electronic health records based on natural language processing," *Appl. Sci.*, vol. 11, no. 19, p. 8812, Sep. 2021.
[50] J. Park, S. C. You, E. Jeong, C. Weng, D. Park, J. Roh, D. Y. Lee, J. Y. Cheong, J. W. Choi, M. Kang, and R. W. Park, "A framework (SOCRATex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: Development and usability study," *JMIR Med. Informat.*, vol. 9, no. 3, Mar. 2021, Art. no. e23983.
[51] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[52] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," 2019, *arXiv:1909.10351*.
[53] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3820–3831, Dec. 2021.
[54] L. Zhao, X. Qian, Y. Guo, J. Song, J. Hou, and J. Gong, "MSKD: Structured knowledge distillation for efficient medical image segmentation," *Comput. Biol. Med.*, vol. 164, Sep. 2023, Art. no. 107284.
[55] L. G. Nateras, F. Dernoncourt, and T. Nguyen, "Hybrid knowledge transfer for improved cross-lingual event detection via hierarchical sample selection," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2023, pp. 5414–5427.
[56] M. Berglund and B. van der Merwe, "Formalizing BPE tokenization," 2023, *arXiv:2309.08715*.



SANGYEON CHO (Graduate Student Member, IEEE) received the B.S. degree in software engineering from Kookmin University, Seoul, Republic of Korea, in 2023. He is currently pursuing the M.S. degree with Chung-Ang University. His research interests include natural language processing, representation learning, and various multimodal learning.



JANGYEONG JEON (Member, IEEE) received the B.S. degree from the Department of Big Data Statistics, Cheongju University, Cheongju, Republic of Korea, in 2022. He is currently pursuing the M.S. degree in artificial intelligence with Chung-Ang University, Seoul, South Korea. His main research interests include natural and various multimodal large language models.



DONGJOON LEE received the B.S. degree in IT media engineering from Kyonggi University, Suwon, South Korea, in 2022. He is currently pursuing the M.S. degree in artificial intelligence with Chung-Ang University, Seoul, South Korea. His main research interests include deep learning-based survival analysis and explainable AI for time-series classification.



CHANGHEE LEE (Member, IEEE) received the B.S. and M.S. degrees from Korea University, Seoul, South Korea, in 2007 and 2011, respectively, and the Ph.D. degree from the University of California at Los Angeles, USA, in 2021. Since 2021, he has been an Assistant Professor with the Department of Artificial Intelligence, Chung-Ang University, Seoul. Since 2024, he has been an Assistant Professor with the Department of Artificial Intelligence, Korea University. His research interests include integrating multiple modalities, building advanced time-series models and causal inference models, discovering scientific knowledge from data, and interpreting "black-box" machine learning methods.



JUNYEONG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from KAIST, Republic of Korea, in 2015, 2017, and 2021, respectively. He was a Postdoctoral Research Associate with the Artificial Intelligence and Machine Learning Laboratory, School of Electrical Engineering, KAIST. He is currently an Assistant Professor with the Department of AI, Chung-Ang University, Republic of Korea. His research interests include visual-language reasoning, visual question answering, and various video-based problems.