

Detecting Redundant Health Survey Questions: Using Language-agnostic BERT Sentence Embedding (LaBSE)

Kang Sunghoon, Hyeoneui Kim, Hyewon Park, Ricky Taira

Submitted to: JMIR Medical Informatics
on: January 24, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 37

 Multimedia Appendixes 38

 Multimedia Appendix 1..... 38



Detecting Redundant Health Survey Questions: Using Language-agnostic BERT Sentence Embedding (LaBSE)

Kang Sunghoon¹; Hyeoneui Kim¹; Hyewon Park¹; Ricky Taira²

¹ Seoul National University Seoul KR

² University of California, Los Angeles Los Angeles US

Corresponding Author:

Hyeoneui Kim

Seoul National University
103 Daehak-ro, Jongno-gu
Seoul
KR

Abstract

Background: As the importance of PGHD in healthcare and research has increased, efforts to standardize survey-based PGHD to improve its usability and interoperability have been made. Standardization efforts, such as the Patient-Reported Outcomes Measurement Information System (PROMIS) and the NIH Common Data Elements (CDE) repository, provided effective tools for managing and unifying health survey questions. However, Previous methods using ontology-mediated annotation are not only labor-intensive and difficult to scale, but also face challenges in identifying semantic redundancies in survey questions, especially across multiple languages.

Objective: The goal of this work was to compute the semantic similarity among publicly available health survey questions in order to facilitate the standardization of survey-based PGHD.

Methods: We compiled various health survey questions authored in both English and Korean from the NIH CDE Repository, PROMIS, Korean public health agencies, and academic publications. Questions were drawn from various health lifelog domains. A randomized question pairing scheme was used to generate a Semantic Text Similarity (STS) dataset consisting of 1758 question pairs. Similarity scores between each question pair were assigned by two human experts. The tagged dataset was then used to build four classifiers featuring: Bag-of-Words, SBERT with BERT-based embeddings, SBRET with LaBSE embeddings, and GPT-4o. The algorithms were evaluated using traditional contingency statistics.

Results: Among the three algorithms, SBERT-LaBSE demonstrated the highest performance in assessing question similarity across both languages, achieving an Area Under the Receiver Operating Characteristic (ROC) and Precision-Recall Curves of over 0.99. Additionally, it proved effective in identifying cross-lingual semantic similarities.

Conclusions: This study introduces the SBERT-LaBSE algorithm for calculating semantic similarity across two languages, showing it outperforms BERT-based models, GPT-4o model and Bag of Words approach, highlighting its potential to improve semantic interoperability of survey-based PGHD across language barriers.

(JMIR Preprints 24/01/2025:71687)

DOI: <https://doi.org/10.2196/preprints.71687>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all.
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/preprint/71687>, the full manuscript will be made available to all.
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/71687>, the full manuscript will be made available to all.



Original Manuscript

Detecting Redundant Health Survey Questions Using Language-agnostic BERT Sentence Embedding (LaBSE)

Sunghoon Kang, MS¹, Hyeoneui Kim, RN, MPH, PhD^{2,3*}, Hyewon Park, RN, BSN^{2,4}, Ricky Taira, PhD⁵

¹College of Natural Sciences, Seoul National University, Seoul, Republic of Korea

²College of Nursing, Seoul National University, Seoul, Republic of Korea

³The Research Institute of Nursing Science, Seoul National University, Seoul, Republic of Korea

⁴Samsung Medical Center, Seoul, Republic of Korea

⁵The Department of Radiological Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

* Corresponding Author

Hyeoneui Kim, RN, MPH, PhD

103 Daehak-ro, Jongno-gu, Seoul, 03080, Republic of Korea

Email: ifilgood@snu.ac.kr

Phone: +82) 02-740-8487

Keywords: Person-Generated Health Data (PGHD); Semantic Similarity; Language-agnostic BERT Sentence Embedding (LaBSE); Sentence-BERT (SBERT)

Abstract

Background

As the importance of Person-Generated Health Data (PGHD) in healthcare and research has increased, efforts to standardize survey-based PGHD to improve its usability and interoperability have been made. Standardization efforts, such as the Patient-Reported Outcomes Measurement Information System (PROMIS) and the NIH Common Data Elements (CDE) repository, provided effective tools for managing and unifying health survey questions. However, Previous methods using ontology-mediated annotation are not only labor-intensive and difficult to scale, but also face challenges in identifying semantic redundancies in survey questions, especially across multiple languages.

Objective

The goal of this work was to compute the semantic similarity among publicly available health survey questions in order to facilitate the standardization of survey-based PGHD.

Materials and Methods

We compiled various health survey questions authored in both English and Korean from the NIH CDE Repository, PROMIS, Korean public health agencies, and academic publications. Questions were drawn from various health lifelog domains. A randomized question pairing scheme was used to generate a Semantic Text Similarity (STS) dataset consisting of 1758 question pairs. Similarity scores between each question pair were assigned by two human experts. The tagged dataset was then used to build four classifiers featuring: Bag-of-Words, SBERT with BERT-based embeddings, SBRET with LaBSE embeddings, and GPT-4o. The algorithms were evaluated using traditional contingency statistics.

Results

Among the three algorithms, SBERT-LaBSE demonstrated the highest performance in assessing question similarity across both languages, achieving an Area Under the Receiver Operating Characteristic (ROC) and Precision-Recall Curves of over 0.99. Additionally, it proved effective in identifying cross-lingual semantic similarities.

Discussion

The SBERT-LaBSE algorithm excelled at aligning semantically equivalent sentences across both languages but encountered challenges in capturing subtle nuances and maintaining computational efficiency. Future research should focus on testing with larger multilingual datasets and on calibrating and normalizing scores across the health lifelog domains to improve consistency.

Conclusion

This study introduces the SBERT-LaBSE algorithm for calculating semantic similarity across two languages, showing it outperforms BERT-based models, GPT-4o model and Bag of Words approach, highlighting its potential to improve semantic interoperability of survey-based PGHD across language barriers.

Introduction

Person-Generated Health Data (PGHD) is becoming increasingly important in managing individual health. PGHD encompasses health-related information that individuals create and collect outside traditional clinical environments, helping them monitor and manage their well-being [1,2]. Examples of PGHD include biometric data from wearable devices and self-reported information such as patient-reported outcomes (PROs). Given its potential for continuously capturing health insights beyond healthcare settings, there is growing interest in leveraging PGHD to support clinical care [3, 5, 6]. In parallel, PGHD is increasingly explored as a resource for patient-centered outcomes research [4, 7]. However, the effective use of PGHD faces several challenges, including developing robust data management systems, ensuring data security, deploying it seamlessly into clinical workflows, and maintaining high data quality [6-8].

Standardizing survey-based Person-Generated Health Data (PGHD) is a critical step in enabling its broader use [9]. An important aspect of standardization is to identify redundancies in the form of semantic equivalencies. These may arise, since dependent upon the author, the clarity, tone, tense, directness, and formality of the language can be phrased differently for the same purposeful inquiry. For example, emotional symptoms might be captured by questions like, "Do you feel like withdrawing from family or friends?" or "I don't really want to talk to people around me." This variation makes identifying semantically equivalent questions—and thus standardizing survey-based PGHD—a complex task. Efforts such as the Patient-Reported Outcomes Measurement Information System (PROMIS) and the National Institutes of Health (NIH) Common Data Elements (CDE) repository aim to provide standardized health survey questions. PROMIS, a consensus-based item bank designed for managing patient-reported outcomes (PROs), offers standardized measures that are applicable across various diseases and clinical settings [10-12]. These measures have helped healthcare providers across various clinical settings, including pain management [13], orthopedics

[14], and primary care [15], cancer care [16], manage patient symptoms, tailor treatments, and improve communication between patients and clinicians. The NIH CDE repository, through metadata tagging, also plays a key role in standardizing data elements, including health surveys [17,18]. Both PROMIS and the CDE repository are essential for enhancing the interoperability of health data.

In practice, the deployment of PGHD acquisition applications requires survey questions be drawn from these established standardized resources. Data collected using questions outside of these resources still requires additional efforts to achieve standardization. While previous studies have explored ontology-mediated methods to identify semantically equivalent health questions [10, 11], annotating each question with ontology concepts is labor-intensive and lacks scalability as such knowledge sources expand. As a complementary approach, deep learning and transformer-based methods have been applied to semantic textual similarity (STS) tasks in clinical texts, including radiology and pathology reports [19], clinical notes [20–22], and medical question–answer pairs [23]. A range of models has been explored, such as convolutional neural networks [19], and transformer-based architectures like BERT, RoBERTa, and XLNet [20–22], and Siamese network [23]. Despite their promising performance, most of these efforts have been limited to single-language settings, predominantly English [20–22] or Chinese [19, 23]. Consequently, cross-lingual STS remains underexplored, highlighting the need for standardization efforts that promote semantic interoperability across languages.

To address these challenges, we developed SPURT (Standardized PGHD Utilization Resources and Tools). SPURT supports the standardization and reuse of survey-based PGHD by identifying semantically equivalent questions, and facilitating the storage, retrieval, and sharing of this data. Unlike PROMIS and the NIH CDE Repository, SPURT annotates and stores health survey questions in both English and Korean, while also detecting semantically redundant questions. This ensures the

use of consistent question formats whenever possible.

Technically, assessing semantic similarity between texts is a well-established yet widely applied method for managing text resources [24]. However, SPURT faces two unique challenges in its assigned task. First, it must effectively assess semantic similarities within or between two different languages, English and Korean. While multilingual embeddings can be used to address this challenge [25, 26], they often perform less effectively for low-resource languages like Korean compared to high-resource languages such as English [27]. One common solution is to translate low-resource languages into high-resource ones before embedding, but this approach risks losing or distorting the original meaning [28, 29]. The second challenge is ensuring computational efficiency for real-time semantic comparisons between questions. Calculating semantic similarity using large language models (LLMs) like BERT (Bidirectional Encoder Representations from Transformers) is computationally expensive, with a time complexity of $O(N!)$. For example, computing the similarity of approximately 10,000 sentence pairs can take around 65 hours using a V100 GPU [30]. Given that SPURT is designed to be a real-time, reactive data processing tool, achieving reasonable response times is crucial for its functionality.

This study presents the development of a novel algorithm for detecting redundant questions, addressing the challenges outlined above. The algorithm utilizes Sentence-BERT (SBERT), a variant of BERT designed for efficient sentence-level semantic similarity calculations [30], along with Language-agnostic BERT Sentence Embedding (LaBSE) [31] to enhance multilingual capability. Sentence-BERT is a model specifically designed for calculating STS between sentences, LaBSE is an embedding that supports efficient cross-lingual STS by mapping multilingual sentences into a shared embedding space. The SBERT-LaBSE algorithm integrates the strengths of both models and facilitates the identification of semantically equivalent questions across languages.

Methods

Corpus Description: The Semantic Textual Similarity (STS) dataset

A STS dataset contains text pairs along with predefined similarity scores that quantify their semantic closeness [32-36]. This study developed an STS dataset to fine-tune pre-trained language models and evaluate our algorithms' performance in determining the semantic similarity between health-related questions.

We collected English and Korean questions from self-reported questionnaires covering the five health lifelog domains including diet, physical activity, living environment, stress management, and sleep. English questions (N=1,222) were sourced from the NIH CDE Repository, PROMIS, and academic publications, while Korean questions (N=963) were gathered from online resources provided by public health agencies and hospitals in Korea [17, 37-40].

To build the STS dataset, we began by randomly selecting five seed questions from each of the five health lifelog domains in Korean, resulting in 25 seed questions. For each question, correspondingly similar questions for Korean were identified, resulting in 25 similar seed questions for each language. This correspondence of seed questions was performed in order to minimize the effects of semantic complexity on algorithm performance. We then randomly selected 30 comparison questions for each seed question, which yielded a total of 1500 question pairs (750 in each language).

The gold standard for semantic similarity between the question pairs was determined by two researchers with nursing backgrounds who independently scored the similarity of each question pair, following a standardized scoring protocol (Table 1). The agreement between the researchers, as

measured by Cohen's kappa, varied by the health lifelog domains: 0.91 for diet, 0.72 for living environment, 0.83 for physical activity, 0.86 for sleep, and 1.0 for stress management, with an average Cohen's kappa of 0.86 across all the health lifelog domains.

Upon completion of this annotation process, we observed that the initial distribution of similarity scores was imbalanced, skewed heavily toward lower similarity scores. Only 2.3% of pairs received a score of 4, and 4.9% received a score of 3. To address this imbalance, we supplemented the dataset with an additional 117 English and 142 Korean question pairs from other sources, chosen to increase the frequency of semantically similar (i.e., higher scores) samples in the evaluation STS dataset. These additions brought the final evaluation set to include 820 question pairs (410 in each language), with the following distribution: 12.2%, 30.5%, 26.8%, and 30.5% for scores 4, 3, 2, and 1 respectively.

Table 1. Scoring protocol for semantic similarity.

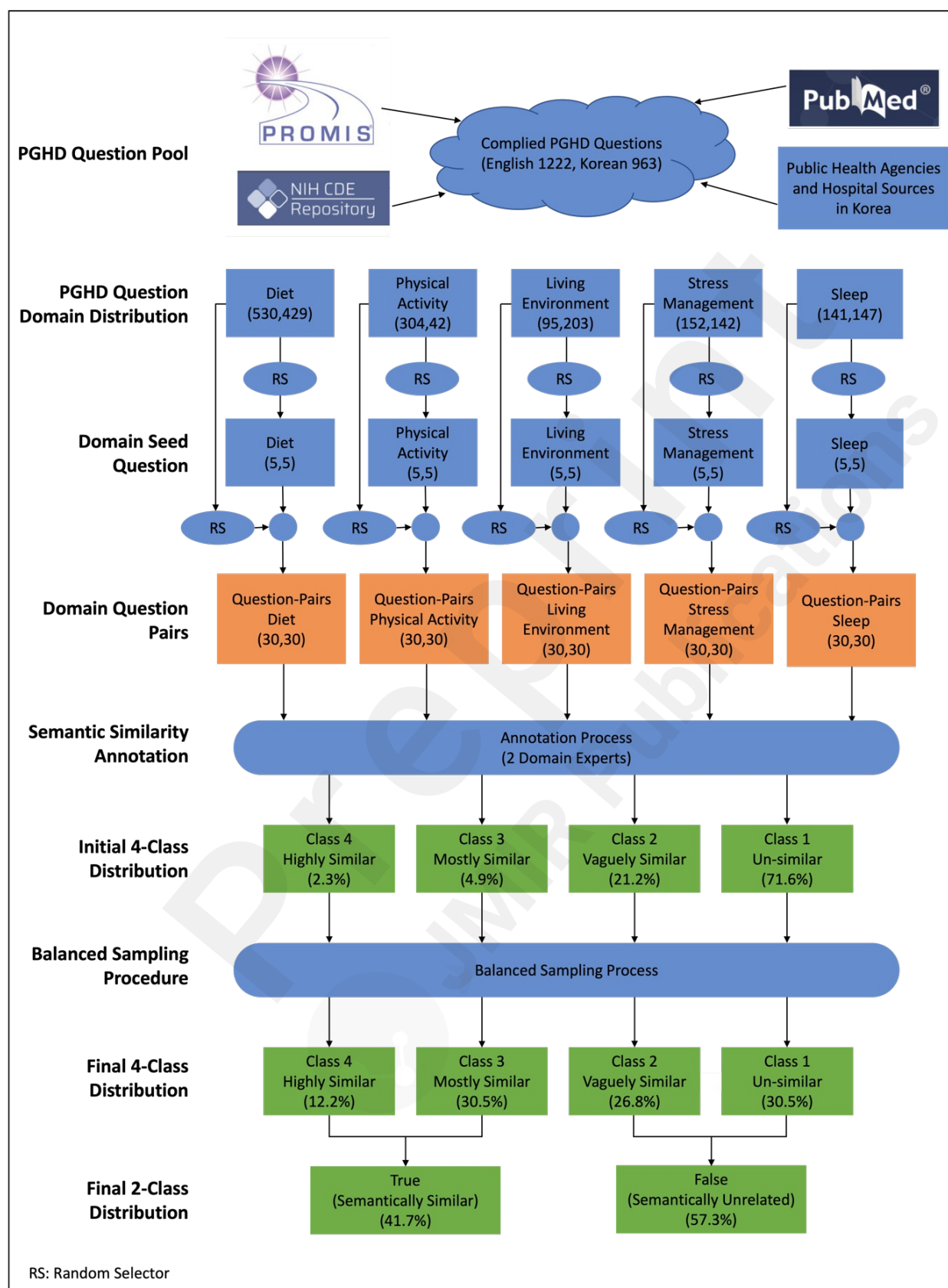
Score	Scoring Protocol	Examples
		Seed question: <i>In the past month, have you ever had chest pain when you were not performing any physical activity?</i>
4	Minor differences in word choice from the seed question, but takes the same form of response	In the past month, have you had chest pain when you were not doing physical activity?
3	Share the same key topics, though some details may be added, altered, or omitted compared to the seed question.	Do you feel pain in your chest when you do physical activity?
2	The key topics are similar but more specific or general than that of the seed question.	Has your doctor ever said that you have a heart condition and that you should only perform physical activity recommended by

a doctor?

- | | | |
|---|--|---|
| 1 | Does not share the core topic from the seed question or belongs to a completely different health lifelog domain. | Have you done general conditioning exercises in the past 4 weeks? |
|---|--|---|
-

Using a similar procedure as described above, we compiled a second English STS dataset for the purpose of fine-tuning our pre-trained language models (see below). This fine-tuning dataset included 938 annotated English question pairs. The fine-tuning set had a distribution of 6.2% scoring 4, 14.0% scoring 3, 23.5% scoring 2, and 56.4% scoring 1.

In total, the STS dataset consisted of 1,758 question pairs, broken down into 820 for evaluation testing (410 English and 410 Korean) and 938 in English for classifier model refinement (see Multimedia Appendix 1, Supplement A). The process of constructing the STS dataset is illustrated in Figure 1.

Figure 1. The process of preparing the STS dataset for fine-tuning and evaluation.

Semantic similarity calculation algorithms

We developed 4 classifiers to compare their performance capability for distinguishing the binary task of semantic similarity between STS question pairs. These included: i) a bag-of-words (BoW) model; ii) a Sentence-BERT with BERT-based embeddings (SBERT-BERT); and iii) a language-agnostic BERT Sentence Embedding model (LaBSE); iv) a GPT-4o Model (GPT-4o). Among these, the SBERT-BERT algorithm was included to serve as a translation-dependent baseline, enabling comparison with multilingual models such as SBERT-BERT and GPT-4o. Model fine-tuning and algorithm development were carried out using Python (version 3.11).

a. Bag of Words (BoW) classifier

The BoW algorithm, a traditional language model that represents sentences by word frequency, serves as the baseline [41]. The BoW model's vocabulary was derived from the STS dataset, comprising 1,349 unique word forms after stop-word removal and lemmatization. Each sentence was represented as a 1,349-dimensional vector based on the vocabulary. Cosine similarity was used to calculate the semantic distance of question pairs. For Korean questions, translation into English was performed using the Google Translator API prior to similarity calculation [42].

b. The SBERT-BERT algorithm

The SBERT-BERT large language model was derived from the pre-trained BERT-base model, which has 12 layers, a 768-dimensional hidden layer, 12 attention heads, and 110 million parameters [30]. SBERT-BERT supports only English. We fine-tuned the pre-trained SBERT-BERT model to optimize its performance for identifying semantic equivalency among health questions using the 938 English question pairs described above. The fine-tuning was performed with a batch size of 32, 8 epochs, and a learning rate of 2e-5, which were deemed optimal after testing various configurations. The AdamW optimizer was used for model optimization [43]. The fine-tuned SBERT-BERT algorithm was then

evaluated using the test STS dataset of 410 English question pairs and 410 Korean question pairs. As previously stated, the Korean questions were translated into English using the Google Translator API, in order to execute the evaluation.

c. The SBERT-LaBSE algorithm

The SBERT-LaBSE algorithm differs from SBERT-BERT in that it supports multiple languages within a single embedding space [31]. The pre-trained SBERT-LaBSE model was derived from the LaBSE model, which also consists of 12 layers, a 768-dimensional hidden layer, 12 attention heads, and 110 million parameters [31]. Fine-tuning was performed in the same manner as for SBERT-BERT. Unlike the other models, SBERT-LaBSE can assess the semantic similarity of English and Korean questions without requiring translation.

d. The GPT-4o algorithm

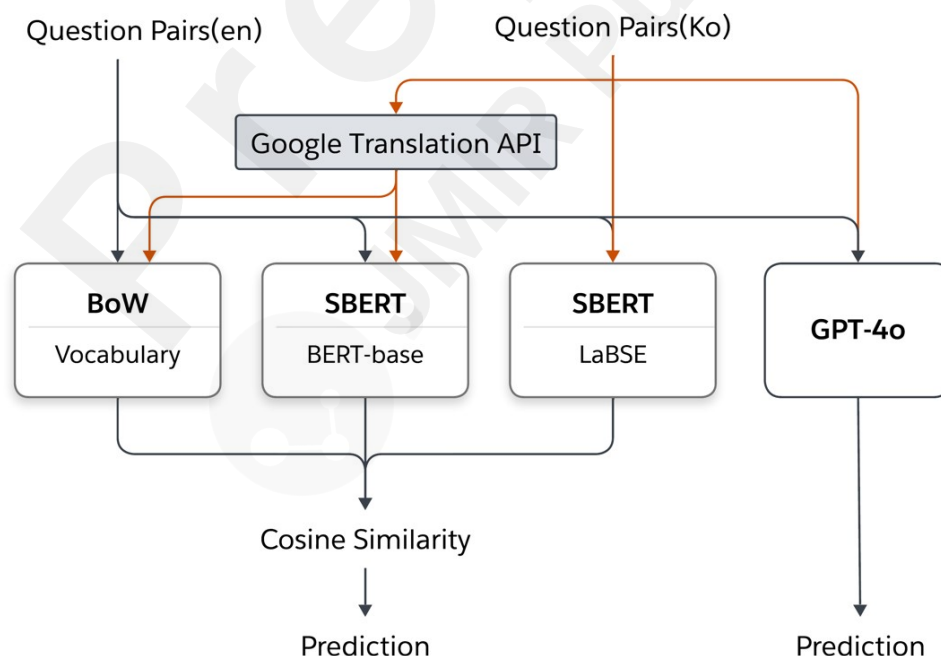
The GPT-4o model, a state-of-the-art LLM, is designed to understand and generate text in multiple languages, including English and Korean [44]. Unlike the SBERT-BERT and SBERT-LaBSE which rely on fixed embeddings for similarity calculation, the GPT-4o operates as a generative model that dynamically evaluates semantic similarity based on contextual understanding. However, in this study, we utilized GPT-4o in a deterministic manner to predict the score of sentence pairs. Each sentence pair was presented with a specific instruction asking to evaluate the score on a scale from 1 to 4 (see Multimedia Appendix 1, Supplement B). Fine-tuning of the GPT-4o model was conducted using the fine-tuning API from OpenAI platform [45].

Performance evaluation

The performance of the similarity calculation algorithms was evaluated as a binary classification problem to simplify interpretation. The 4-point ordinal similarity scores from the STS dataset were converted into binary labels, where scores of 3 and 4 were categorized as 'similar' and scores of 1 and 2 as 'dissimilar.'

Optimal thresholds for predicting similarity were determined for the continuous similarity scores, which ranged from -1 to 1. Precision, recall, and F1-scores were calculated to assess algorithm performance, and the Area Under the Curve (AUC) for both the Receiver Operating Characteristics (ROC) and Precision-Recall (PR) curves were examined. The processes used by the three algorithms to calculate similarity are illustrated in Figure 2.

Figure 2. Similarity calculation with the four algorithms.



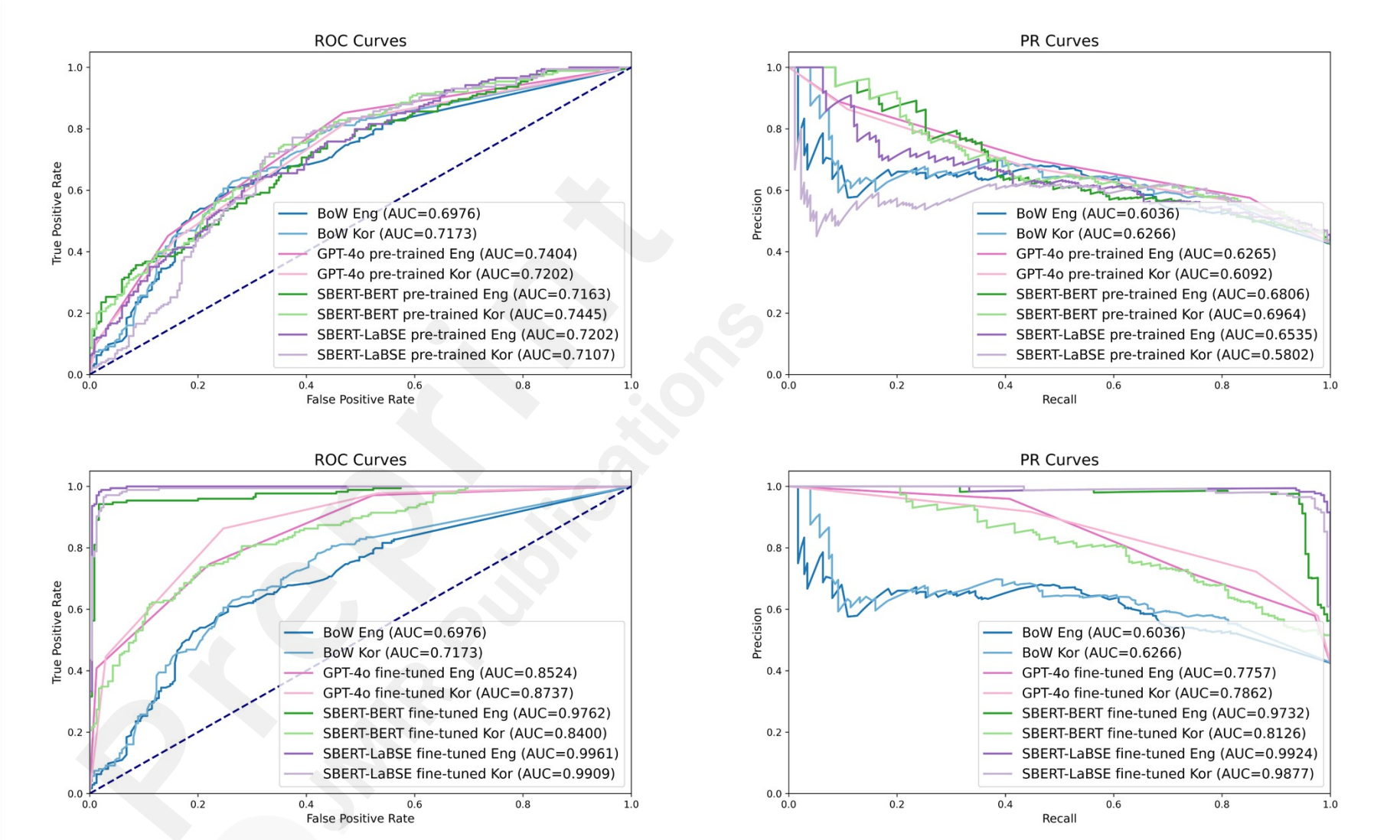
Result

The performance of the three models for classifying similar vs. unsimilar question pairs when aggregating the five health lifelog domains is summarized in Table 2 and Figure 3. In the zero-shot trials (i.e., without the model refining stage), there were minimal differences in performance among the three algorithms for both English and Korean questions. All algorithms exhibited higher recall than precision in both languages. After fine-tuning, the SBERT-BERT algorithm showed a substantial improvement, particularly for English questions, in which the F1 score was raised from 0.65 to 0.96. For Korean questions, the improvement was moderate, with F1 score progressing from 0.68 to 0.73. In contrast, SBERT-LaBSE demonstrated significant improvements for both languages post fine-tuning. For English, F1 scores were raised from 0.66 to 0.98 while for Korean, F1 scores increased from 0.68 to 0.98. Fine tuning for both the SBERT-BERT and the SBERT-LaBSE models resulted in noticeable balanced performance between recall and precision. Similarly, GPT-4o exhibited improved performance following fine-tuning, with its F1 scores increasing from 0.69 to 0.79 for English and from 0.67 to 0.79 for Korean. However, the degree of improvement was smaller compared to that observed in the SBERT models.

Language	Performance Metrics	BoW	GPT-4o pre-trained	GPT-4o fine-tuned	SBERT with pre-trained		SBERT with fine-tuned	
					BERT-base	LaBSE	BERT-base	LaBSE
English question pairs (N=410)	Accuracy	0.6112	0.6683	0.8463	0.6308	0.5917	0.9702	0.9853
	Precision	0.5279	0.5753	0.9590	0.5451	0.5111	0.9668	0.9818
	Recall	0.8161	0.8514	0.6686	0.7989	0.9253	0.9632	0.9839
	F1	0.6411	0.6866	0.7879	0.6480	0.6585	0.9649	0.9828
Korean question pairs (N=410)	Accuracy	0.6610	0.6512	0.8488	0.6659	0.6878	0.7576	0.9839
	Precision	0.5732	0.5620	0.9520	0.5760	0.6054	0.6929	0.9818
	Recall	0.8057	0.8286	0.6800	0.8229	0.7714	0.7817	0.9806
	F1	0.6698	0.6697	0.7933	0.6776	0.6784	0.7332	0.9812

Table 2. Performance Metrics for the three algorithms, combining the health lifelog domains. BoW: Bag of Words

Figure 3. ROC and PR curves for pre-trained and fine-tuned embeddings on English and Korean questions, combining the health lifelog domains



	Performance Metrics	English question pairs (N=410)						Korean question pairs (N=410)					
		DL (N=80)	HLE (N=80)	PA (N=80)	Sleep (N=85)	Stress (N=85)	All	DL (N=80)	HLE (N=80)	PA (N=80)	Sleep (N=85)	Stress (N=85)	All
BoW	Accuracy	0.7215	0.7250	0.6625	0.4118	0.7176	0.6112	0.7250	0.8125	0.7250	0.5765	0.5882	0.6610
	Precision	0.7000	0.6383	0.5952	0.4118	0.6279	0.5279	0.6585	0.7941	0.6275	0.0000	0.5000	0.5732
	Recall	0.6176	0.8571	0.7143	1.0000	0.7714	0.8161	0.7714	0.7714	0.9143	0.0000	0.8286	0.8057
	F1	0.6563	0.7317	0.6494	0.5833	0.6923	0.6411	0.7105	0.7826	0.7442	0.0000	0.6237	0.6698
	ROC AUC	0.7297	0.7457	0.6810	0.5820	0.7611	0.6976	0.7667	0.7937	0.7933	0.5937	0.6609	0.7174
	PR AUC	0.7250	0.6718	0.6301	0.5025	0.6834	0.6036	0.7394	0.7373	0.7519	0.4498	0.5985	0.6267
GPT-4o Fine-tuned	Accuracy	0.9367	0.9750	0.9625	0.9765	0.9765	0.7873	0.8875	0.9375	0.7750	0.8000	0.8471	0.8293
	Precision	0.8919	0.9714	0.9444	0.9459	0.9459	0.7403	0.9643	1.0000	0.8696	0.9500	0.9583	0.7586
	Recall	0.9706	0.9714	0.9714	1.0000	1.0000	0.7701	0.7714	0.8571	0.5714	0.5429	0.6571	0.8800
	F1	0.9296	0.9714	0.9577	0.9722	0.9722	0.7549	0.8571	0.9231	0.6897	0.6909	0.7797	0.8148
	ROC AUC	0.9598	0.9838	0.9727	0.9863	0.9757	0.8524	0.9340	0.9444	0.8295	0.8823	0.8769	0.8737
	PR AUC	0.9138	0.9672	0.9418	0.9629	0.9354	0.7757	0.8969	0.9344	0.7510	0.7969	0.8182	0.7862
SBERT with Fine-tuned BERT-base	Accuracy	0.9646	0.9625	0.9800	0.9906	0.9835	0.9702	0.8525	0.8125	0.7025	0.7200	0.7882	0.7576
	Precision	0.9650	0.9502	0.9784	0.9836	0.9830	0.9668	0.8037	0.7391	0.6175	0.6036	0.7108	0.6929
	Recall	0.9529	0.9657	0.9771	0.9943	0.9771	0.9632	0.8800	0.8914	0.8629	0.9543	0.8229	0.7817
	F1	0.9585	0.9571	0.9770	0.9887	0.9799	0.9649	0.8384	0.8062	0.7176	0.7376	0.7622	0.7332
	ROC AUC	0.9859	0.9698	0.9923	0.9929	0.9936	0.9867	0.9125	0.8563	0.7901	0.8411	0.8462	0.8412
	PR AUC	0.9858	0.9480	0.9925	0.9870	0.9918	0.9800	0.9008	0.7969	0.7640	0.8109	0.8244	0.8134
SBERT with Fine-tuned LaBSE	Accuracy	0.9848	0.9900	0.9875	0.9906	0.9906	0.9853	0.9775	0.9975	0.9850	0.9859	0.9835	0.9839
	Precision	0.9716	0.9889	0.9728	0.9944	0.9889	0.9818	0.9719	0.9944	0.9836	0.9775	0.9886	0.9818
	Recall	0.9941	0.9886	1.0000	0.9829	0.9886	0.9839	0.9771	1.0000	0.9829	0.9886	0.9714	0.9806
	F1	0.9826	0.9885	0.9861	0.9884	0.9887	0.9828	0.9743	0.9972	0.9828	0.9829	0.9797	0.9812
	ROC AUC	0.9965	0.9929	0.9987	0.9989	0.9979	0.9968	0.9893	0.9976	0.9962	0.9971	0.9930	0.9951
	PR AUC	0.9964	0.9901	0.9984	0.9985	0.9975	0.9960	0.9872	0.9958	0.9947	0.9957	0.9927	0.9934

Table 3. Performance Metrics of the SBERT based algorithms with the fine-tuned BERT and LaBSE models by the health lifelog domains. DL: Dietary Lifestyle; HLE: Human Living Environment; PA: Physical Activity.

Table 3 presents the performance of the two SBERT algorithms across the five health lifelog domains. For all the health lifelog domains, the fine-tuned SBERT-BERT and SBERT-LaBSE algorithms demonstrated high performance on English questions, with ROC AUC and PR AUC values exceeding 0.95 and approaching 0.99. However, the SBERT-BERT algorithm struggled with the English-translated Korean questions, particularly in the Physical Activity domain. In contrast, the SBERT-LaBSE algorithm consistently delivered strong performance across all the health lifelog domains even for Korean questions.

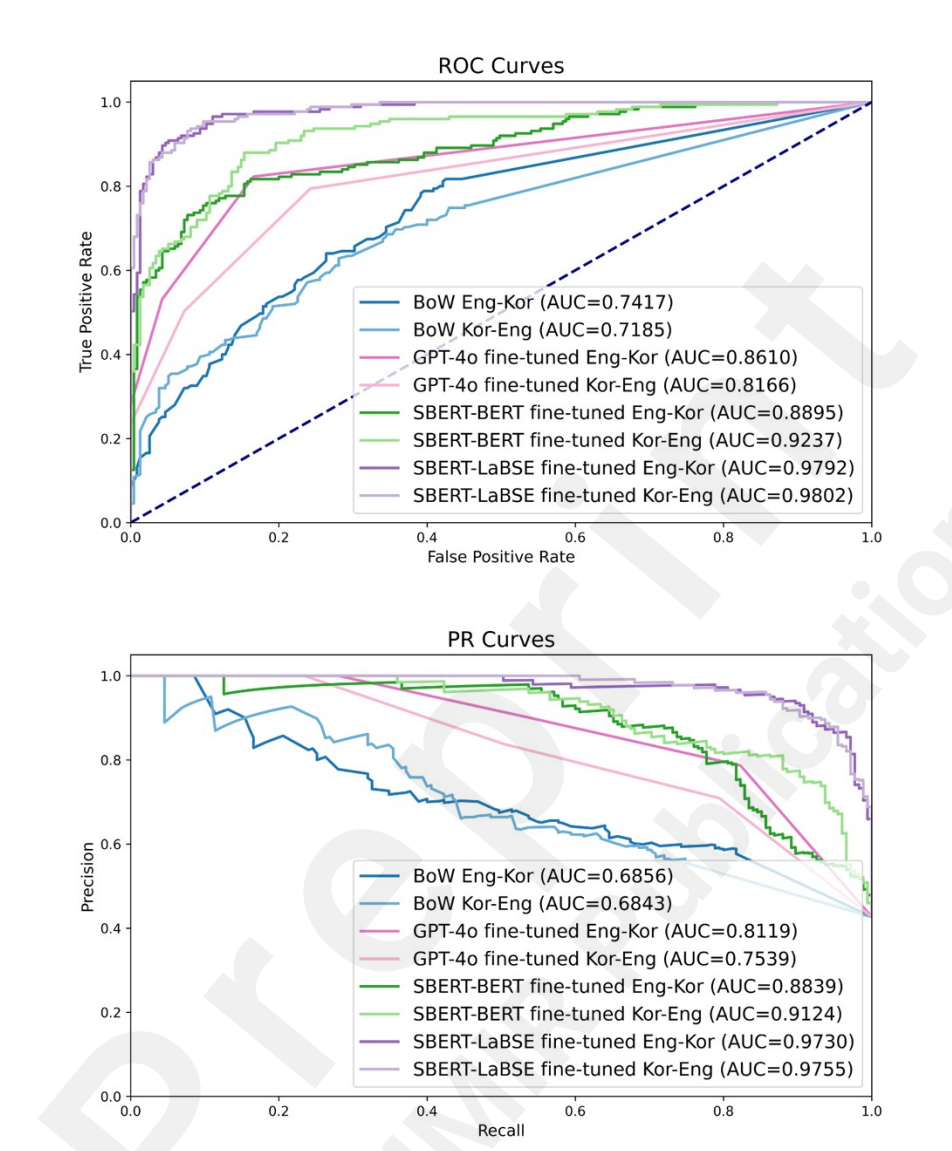
Table 4 presents the optimal cut-off values for the three algorithms. The pre-trained SBERT-BERT and SBERT-LaBSE models showed considerable variation in cut-off values across the five health lifelog domains. However, after fine-tuning, these variations decreased, indicating that fine-tuning helped stabilize the algorithms. Despite this improvement, the SBERT-LaBSE algorithm still exhibited more variability in cut-off values across the health lifelog domains compared to SBERT-BERT, suggesting that further calibration may be required for SBERT-LaBSE. Supplement C of Multimedia Appendix 1 provides example question pairs from each health lifelog domain, along with the similarity scores assigned by human reviewers and predicted by the three algorithms.

Table 4. Optimal cut-off for algorithms on BoW and pre-trained and fine-tuned SBERT-BERT and SBERT-LaBSE in each health lifelog domain.

Language	Health lifelog Domain	Bag of Words	SBERT with pre-trained		SBERT with fine-tuned	
			BERT-base	LaBSE	BERT-base	LaBSE
English question pairs (N=410)	DL	0.2887	0.6274	0.5359	0.6349	0.6262
	HLE	0.1291	0.5369	0.3965	0.6151	0.6425
	PA	0.3162	0.3667	0.4822	0.6304	0.6202
	Sleep	0.0000	0.6790	0.2456	0.6617	0.6574
	Stress	0.1054	0.5817	0.3807	0.6359	0.5958
	All	0.1291	0.5816	0.3796	0.6278	0.6091
Korean question pairs (N=410)	DL	0.2887	0.5990	0.3103	0.5639	0.6568
	HLE	0.2582	0.5475	0.5603	0.5639	0.7138
	PA	0.1491	0.4778	0.6004	0.5639	0.6741
	Sleep	0.9354	0.4837	0.9215	0.5639	0.6849
	Stress	0.1091	0.6647	0.4481	0.5639	0.6586
	All	0.1336	0.5320	0.5753	0.5639	0.6531

Figure 4 illustrates that SBERT-LaBSE effectively determined semantic similarities between the two languages, with slightly better performance in identifying the semantic similarities of English questions relative to the Korean seed questions. The full results of the cross-language semantic similarity analysis are provided in Supplement D of Multimedia Appendix 1.

Figure 4. Performance of the cross-language semantic similarity determination.



Discussion

Principal Results

This study demonstrates the utility of large language models for the purpose of determining semantic similarities among health questions in order to facilitate the standardization of survey-based health data. The results indicate that the fine-tuned SBERT algorithms were significantly more effective than the traditional BoW approach in identifying semantic similarities for both English and Korean questions. Furthermore, the SBERT-LaBSE algorithm demonstrated superior performance, particularly for Korean questions, suggesting that it is a more effective method than the SBERT-BERT algorithm, which relies on English translation, for assessing semantic similarity in non-English texts. Notably, the SBERT-LaBSE algorithm outperformed the GPT-4o algorithm, particularly in Korean. While it is possible that the full potential of the GPT-4o algorithm was not realized, the results clearly show that for the specific task examined in this study, the fine-tuned SBERT algorithms achieved better performance compared to GPT-4o, with significantly lower computational costs [46].

The SBERT-LaBSE algorithm's success with Korean questions can be attributed to its structural design and the limitations of language translation. Structurally, LaBSE aligns semantically equivalent words or sentences from different languages into a unified embedding space, preserving semantic consistency across languages. This allows for more accurate semantic similarity assessments. In contrast, the SBERT-BERT algorithm's lower performance with Korean questions may be due to meaning loss or distortion during translation, which disrupts semantic comparisons between languages [28,29]. While previous studies have noted that LaBSE may struggle with subtle, sentence-level nuances, limiting its

performance in fine-grained similarity tasks [47], our study found that the SBERT-LaBSE algorithm effectively captured the meanings in both English and Korean sentences, outperforming the SBERT-BERT model. However, this finding should be validated with a larger, more diverse dataset that includes a broader range of syntactic features.

Limitations

When implemented in the SPURT system with 1,835 questions in the comparison space, the SBERT-LaBSE algorithm evaluated the similarity of a new question in just 0.03 seconds. This was achieved on a Naver Cloud Platform server with 8GB RAM and no GPU [48]. Despite its impressive performance, LaBSE's 440 million parameters—four times that of BERT-base—make it a resource-intensive option, potentially increasing costs for complex tasks. This resource demand may limit its applicability on resource-constrained devices such as mobile platforms [49]. To address these limitations, future work will explore techniques such as distillation [50] and the use of small language models (SLMs) [51], with the goal of reducing model size while maintaining performance.

This study has some limitations. First, the cut-off values for the similarity scores were not uniformly calibrated across the five health lifelog domains, leading to inconsistencies in how similarity scores were interpreted. For example, the SBERT-LaBSE algorithm assigned a similarity score of 0.7 to the dietary question pair "I've binge eaten" and "Do you ever overeat?" and identified them as similar. On the other hand, the algorithm correctly identified the human living environment questions "Have you moved in the past 5 years?" and "In the last 5 years, the number of people in this community has?" dissimilar while assigning the same similarity score of 0.7 to the pair. These inconsistencies may impact the accurate interpretation of similarity scores, highlighting the need for future work to focus on

calibrating and normalizing scores across the health lifelog domains to ensure greater consistency.

Second, the current evaluation was conducted on a relatively small set of English and Korean question pairs. Future studies should explore the feasibility of applying the SBERT-LaBSE algorithm to a broader range of sentence types from diverse domains. Additionally, by incorporating texts from more diverse languages, future research can investigate the algorithm's potential to overcome language barriers and facilitate semantic interoperability.

Comparison with Prior Work

Previous methods that relied on metadata tagging [17, 18] and ontology-mediated annotation [10, 11] were effective in providing structured mappings between concepts, facilitating interoperability. However, they struggled with comparing the meanings of survey questions composed in multiple languages and addressing semantically redundant questions. This study leverages fine-tuned LLMs such as SBERT-BERT and SBERT-LaBSE to assess semantic similarity. Particularly, the fine-tuned SBERT-LaBSE algorithm demonstrates the potential to enhance semantic interoperability by capturing semantic similarities across multiple languages with high performance.

Conclusion

This study showcases the potential of large language models for identifying semantic redundancy in survey-based Person-Generated Health Data (PGHD) collections. Specifically, the SBERT-LaBSE algorithm excelled in classifying semantic similarity across diverse question formats in two languages. Our findings demonstrate that SBERT-LaBSE outperforms traditional BERT-based algorithm, the GPT-4o algorithm, and the conventional

bag-of-words (BoW) approach in both languages, highlighting its capacity to improve semantic interoperability of PGHD across language barriers.



Acknowledgments

We are thankful to our colleagues, Eunyeong Lim and Jeongha Kim, for their help with the STS dataset preparation. This study was supported in part by a grant from the National Research Foundation of Korea (grant number:2022R1A2C201136011).

Conflicts of interests

None declared.

References

1. Shapiro M, Johnston D, Wald J, Mon D. Patient-generated health data. White paper. RTI International; 2012.
2. Office of the National Coordinator for Health Information Technology (ONC). What are patient-generated health data? Accessed July 22, 2024. Available from: <https://www.healthit.gov/topic/scientific-initiatives/pcor/patient-generated-health-data-pghd>.
3. Isind AS, Lindroth T, Lundin J, Steineck G. Shift in translations: Data work with patient-generated health data in clinical practice. *Health Inform J*. 2019;25(3):577-86.
4. Chung AE, Sandler RS, Long MD, et al. Harnessing person-generated health data to accelerate patient-centered outcomes research: the Crohn's and Colitis Foundation of America PCORnet Patient Powered Research Network (CCFA Partners). *J Am Med Inform Assoc*. 2016;23(3):485-490.
5. Sanger PC, Hartzler A, Lordon RJ, et al. A patient-centered system in a provider-centered world: challenges of incorporating post-discharge wound data into practice. *J Am Med Inform Assoc*. 2016;23(3):514-525.
6. Adler-Milstein J, Nong P. Early experiences with patient generated health data: health system and patient perspectives. *J Am Med Inform Assoc*. 2019;26(10):952-959.
7. Austin E, Lee JR, Amtmann D, et al. Use of patient-generated health data across healthcare settings: implications for health systems. *JAMIA Open*. 2019;3(1):70-76.
8. Abdolkhani R, Gray K, Borda A, DeSouza R. Patient-generated health data management and quality challenges in remote patient monitoring. *JAMIA Open*. 2019 Sep 20;2(4):471-478.

9. Hussein R, Crutzen R, Gutenberg J, Kulnik ST, Sareban M, Niebauer J. Patient-Generated Health Data (PGHD) Interoperability: An Integrative Perspective. *Stud Health Technol Inform*. 2021;281:228-232.
10. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Med Care*. 2007;45.
11. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*. 2010;63(11):1179-1194.
12. Pilkonis PA, Choi SW, Reise SP, et al. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment*. 2011;18(3):263-283.
13. Amtmann D, Cook KF, Jensen MP, et al. Development of a PROMIS item bank to measure pain interference. *Pain*. 2010;150(1):173-182.
14. Brodke DJ, Saltzman CL, Brodke DS. PROMIS for Orthopaedic Outcomes Measurement. *J Am Acad Orthop Surg*. 2016;24(11):744-749.
15. Kroenke K, Talib TL, Stump TE, et al. Incorporating PROMIS symptom measures into primary care practice—a randomized clinical trial. *J Gen Intern Med*. 2018;33(8):1245-1252.
16. Wagner LI, Schink J, Bass M, et al. Bringing PROMIS to practice: Brief and precise symptom screening in ambulatory cancer care. *Cancer*. 2015.
17. National Institutes of Health. Common Data Elements Repository. Accessed October 14, 2024. Available from: <https://cde.nlm.nih.gov/home>.
18. National Institutes of Health. NIH Strategic Plan for Data Science. 2018.

19. Zheng T, Gao Y, Wang F, Zhang Y, Lin H, Chen Q. Detection of medical text semantic similarity based on convolutional neural network. *BMC Med Inform Decis Mak.* 2019;19(1):156.
20. Yang X, Bian J, Hogan WR, Wu Y. Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models. *JMIR Med Inform.* 2020;8(11):e19735.
21. Mahajan D, Joshi A, Lehman A, Zheng K, Li Z, Wang Y, et al. Identification of semantically similar sentences in clinical notes: iterative intermediate training using multi-task learning. *JMIR Med Inform.* 2020;8(11):e22508.
22. Ormerod M, Martínez Del Rincón J, Devereux B. Predicting semantic similarity between clinical sentence pairs using transformer models: evaluation and representational analysis. *JMIR Med Inform.* 2021;9(5):e23099.
23. Li Q, He S. Similarity matching of medical question based on Siamese network. *BMC Med Inform Decis Mak.* 2023;23(1):55.
24. Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. *AAAI Conference on Artificial Intelligence*; 2006.
25. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the North American Chapter of the Association for Computational Linguistics*; 2019.
26. Artetxe M, Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans Assoc Comput Linguist.* 2019;7:597-610.
27. Miao Z, Wu Q, Zhao K, Wu Z, Tsuruoka Y. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. *Proceedings of the Association for Computational Linguistics: NAACL*; 2024.
28. Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine

- translation. *International Conference on Learning Representations*; 2018.
29. Tiyaamorn N, Kajiwaru T, Arase Y, Onizuka M. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*; 2021.
 30. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; 2019.
 31. Feng F, Yang Y, Cer DM, Arivazhagan N, Wang W. Language-agnostic BERT Sentence Embedding. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; 2022.
 32. Agirre E, Cer DM, Diab MT, Gonzalez-Agirre A. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. *Proceedings of the International Workshop on Semantic Evaluation*; 2012.
 33. Gao T, Yao X, Chen D. SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; 2021.
 34. Ham J, Choe YJ, Park K, Choi I, Soh H. KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding. *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP*; 2020.
 35. Mutinda FW, Yada S, Wakamiya S, Aramaki E. Semantic textual similarity in Japanese clinical domain texts using BERT. *Methods Inf Med*. 2021;60.
 36. Zhang N, Chen M, Bi Z, Liang X, Li L, et al. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*; 2022.
 37. HealthMeasures: Transforming how health is measured. Accessed October 14, 2024.

Available from: <https://www.healthmeasures.net>.

38. Ministry of Food and Drug Safety. Accessed October 14, 2024. Available from: <https://www.mfds.go.kr/index.do>.
39. National Cancer Information Center. Accessed October 14, 2024. Available from: <https://www.cancer.go.kr/lay1/S1T226C228/contents.do>.
40. Seoul Asan Medical Center. Health Screening & Promotion Center. Accessed October 15, 2024. Available from: <https://health.amc.seoul.kr/health/maintain/check.do>.
41. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM*. 1975;18(11):613-620.
42. Googletrans: Python Library that Implement Google Translate API. Accessed October 10, 2024. Available from: <https://pypi.org/project/googletrans-py/>.
43. AdamW PyTorch documentation. Accessed October 10, 2024. Available from: <https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html>.
44. OpenAI. "GPT-4o System Card". 2024. Accessed January 19, 2025. Available from: <https://openai.com/research/gpt-4o-system-card>.
45. OpenAI. Fine-tuning guide. Accessed January 19, 2025. Available from: <https://platform.openai.com/docs/guides/fine-tuning>.
46. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-1901.
47. Wang W, Chen G, Wang H, Han Y, Chen Y. Multilingual sentence transformer as a multilingual word aligner. *Findings of the Association for Computational Linguistics: EMNLP*; 2022.
48. Naver Cloud Platform. Accessed October 14, 2024. Available from: <https://www.ncloud.com/product/compute/server#detail>

49. Mao Z, Nakagawa T. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*; 2023.
50. Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, et al. TinyBERT: distilling BERT for natural language understanding. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*; 2020. p. 4163–74.
51. Wang F, Zhang Z, Zhang X, Wu Z, Mo T, Lu Q, et al. A comprehensive survey of small language models in the era of large language models: techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. *arXiv preprint arXiv:2411.03350*. 2024.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers

BoW: Bag of Words

CDE: Common Data Elements

LaBSE: Language-agnostic BERT Sentence Embedding

LLM: Large Language Model

PGHD: Person-Generated Health Data

PROMIS: Patient-Reported Outcomes Measurement Information System

SBERT: Sentence-BERT

STS: Semantic Textual Similarity

Supplementary Files

Multimedia Appendixes

Comparison of algorithms' performance on the STS dataset.

URL: <http://asset.jmir.pub/assets/f34dce76e1364bdd99d36fc842467e7c.docx>