JDFR Journal of Data Forensics Research
데이터포렌식연구, 1권 1호, 2024년 12월, 109-123면
Journal of Data Forensics Research, Volume 1, Issue 1, December 2024, Pages 109-123.
https://doi.org/10.12972/JDFR.2024.1.1.7
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

설문조사자료의 문항별 문맥 반영을 위한 자연어 임베딩 활용연구

정재경[1], 오미애[2], 신성철[3], 최호식[4]

[1]서울시립대학교 도시빅데이터융합학과 박사재학
[2]한국보건사회연구원 연구위원
[3]크립토랩 데이터사업개발실 이사
[4]서울시립대학교 도시빅데이터융합학과 교수

교신저자: 최호식, choi.hosik@uos.ac.kr

요약

설문조사 응답 데이터는 일반적으로 수치화되어 분석되지만, 각 질문의 응답을 단일 점수로 활용할 경우 질문의 의도나 중요도가 반영된 문맥이 간과될 수 있다. 이를 보완하기 위해, 본 연구에서는 질문과 응답을 자연어 문장으로 구성하고, 거대언어모델을 활용하여 자연어 임베딩을 생성하는 방법론을 탐구한다. 사전 학습된 다양한 거대언어모델을 활용하여 질문-응답 결합 문장의 임베딩 벡터를 비교하고, 문항별 문맥 반영에 가장 적합한 모델을 선정한다. 제안된 방법론을 정신건강 관련 설문조사 데이터에 적용하여 자살사고, 자살계획, 자살시도 대상자의 집단적 특성을 파악한 결과, 기존의 점수 기반 평가 방식과 비교하여 질문-응답 쌍으로부터 생성된 임베딩 벡터가 정신건강 위험군 분류에 더 효과적임을 실증하였다.

주제어

거대언어모델, 임베딩 벡터, XGBoost, 2021년 정신건강실태조사

Received: November 27, 2024
Revised: December 23, 2024
Accepted: December 24, 2024
Published: December 31, 2024

Original Article

Context-Aware Classification Method for Survey Data Using Natural Language Embeddings

Jaekyeong Jung[1], Miae Oh[2], Sungcheol Shin[3], Hosik Choi[4]

[1]Ph.D. Candidate, Department of Urban Big Data Convergence, University of Seoul, Republic of Korea
[2]Research Fellow, Korea Institute for Health and Social Affairs, Republic of Korea
[3]Director, Data Business Development Division, CryptoLab Inc, Republic of Korea
[4]Professor, Department of Urban Big Data Convergence, University of Seoul, Republic of Korea

Corresponding Author: Hosik Choi, choi.hosik@uos.ac.kr

ABSTRACT

Survey response data is typically quantified for analysis; however, when each question is represented as a single score, the context reflecting the intent and importance may be overlooked. To address this, we propose a method that generates natural language embeddings by concatenating questions and responses through large language models (LLMs). By comparing embedding vectors derived from question-response pairs using various pre-trained LLMs, we identify the model most suitable for capturing contextual information. Applying the proposed approach to mental health survey data, we identified the collective characteristics of individuals with suicidal thoughts, plans, and attempts, demonstrating that embedding vectors extracted from question-response pairs are more effective in classifying mental health risk groups compared to traditional score-based evaluation methods.

KEYWORDS

Large language model, embedding vector, XGBoost, National Mental Health Survey of Korea 2021

1. 서론

최근 묻지마범죄, 이상동기 범죄, 무차별 범죄 등이 급증하고 있다. 이러한 범죄와 직간접적인 관계가 있는 정신건강은 단순히 정신병이나 심리적 장애가 없는 상태가 아니라, 개인의 삶에 대한 만족도와 적응력, 사회적 관계와 역할 수행능력 등을 포함하는 포괄적인 개념이다. 정신장애는 개인과 사회적 환경 요인이 상호작용하는 과정에서 악화되거나 지속될 수 있는 상태이며 사회통합을 저해할 수 있다[1]. 지역사회의 정신건강 실상을 파악하고 효과적인 예방과 치료 근거를 제공하기 위해 수집되는 정신건강 데이터를 세밀하게 해석해야 한다.

본 연구의 분석자료인 [2021년 정신건강실태조사 마이크로데이터]는 자연어 형태의 객관식 문항과 주관식 문항으로 이루어져 있다. 응답자는 질문을 읽고 본인의 생각 또는 상황에 가장 적절한 보기를 선택하거나 의견을 작성한다. 일반적으로 설문조사 데이터는 범주화된 보기를 기반으로 정량화하여 해석된다. 이를 통해 응답의 전체적인 분포를 파악하거나 종합적인 점수를 계산하여 응답자의 상태나 수준을 비교할 수 있다. 그러나, '니코틴 사용에 의한 장애', '공포장애 및 기타 불안 장애' 등 6개의 대분류로 구조화된 설문문항들로부터 응답여부의 이항(binary)값을 합산하여 단일화된 수치로 평가하는 방식은 질문이 가지고 있는 문맥(context)과 이에 대한 응답의 문맥을 고려하는데 일정 수준의 한계를 수반한다.

질문이 내포하는 문맥을 가중치로 고려하기 위한 방안으로는 자연어 처리분야에서 활용되는 언어모델로부터 구할 수 있는 문장의 임베딩(embedding) 기법을 활용할 수 있다. 임베딩 벡터는 자연어의 밀집 표현(dense representation)으로 단어들의 동시 출현빈도를 기초로 학습되는 다차원의 실수공간상에 놓여있는 벡터를 의미한다. 트랜스포머(transformer, [2]) 모델 아키텍처를 기반으로 한 거대언어모델은 자연어의 임베딩 표현(embedding representation)을 유연한 방식으로 학습한다. 트랜스포머는 멀티헤드 어텐션(multi-head attention)을 통해 단어 간의 다양한 관계를 병렬적으로 학습하며, 포지셔널 인코딩(positional encoding)을 활용해 문장의 순서와 문법적 구조를 반영한다. 이러한 구조적 특징으로 자연어 문장의 다차원적인 의미와 문맥을 효과적으로 반영한 풍부한 특징 벡터(feature vector)를 생성할 수 있다.

본 연구에서는 설문조사 데이터의 질문에 내포된 문맥으로부터 위험도를 반영할 수 있는 자연어처리 기반 분석 방법을 적용하여 추가정보를 추출하고 이를 활용하는 방안을 모색해보고자 한다. 2절에서는 문장 임베딩 방법론을 살펴보고, 데이터셋과 이를 전처리하는 과정을 소개한다. 또한, 사전학습된(pre-trained) 언어모델을 선정하기 위한 소규모 실험 결과에 기반한 모델 선택 방법을 설명한다. 3절에서는 선정된 사전학습된 언어모델을 활용하여 설문문항을 임베딩하고 t-SNE 군집방법[3,4]을 활용하여 임베딩 방법이 저차원상에서의 구조화된 설문문항군의 군집성을 효과적으로 표현할 수 있음을 설명한다. 또한, 각 설문문항을 2차원의 벡터로 축소한 정보로 정신건강 집단군을 분류하는 실험을 수행한다. 끝으로, 결론과 향후계획에 대해서 서술한다.

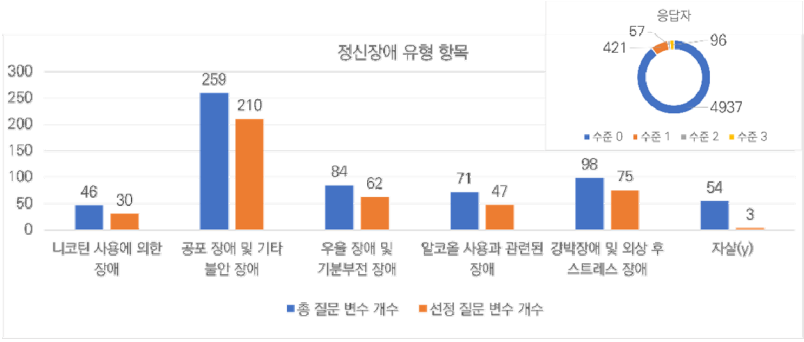
2. 방법론

본 절에서는 정신건강실태조사 마이크로데이터를 소개하고, 사전학습된 언어모델을 활용한 설문지 문항 임베딩 방법론들의 특징을 살펴본다.

2.1. 데이터셋과 전처리

분석 대상은 2021년 정신건강실태조사 마이크로데이터 코드북 변수 중, B(니코틴 사용에 의

한 장애), D(공포 장애 및 기타 불안 장애), E(우울 장애 및 기분부전 장애), J(알코올 사용과 관련된 장애), K(강박장애 및 외상 후 스트레스 장애), S(자살) 항목으로 선정하였다. 집단적인 응답 패턴을 해석하기 위하여 나이와 기간처럼 개인 특성을 묻는 질문은 제외하고 보기가 '예'와 '아니오'의 이항 형태로 이루어진 질문 변수만 택하였다. <Figure 1>은 분석에 사용한 항목별 질문의 개수를 나타낸다. 6가지 항목의 정신장애 유형에 대해 총 424개의 질문을 채택하였다. 전체 응답자는 S(자살) 유형의 생각, 계획, 시도 응답변수를 기준으로 정신건강 위험 수준을 4개 그룹으로 구분하였다.



<Figure 1> 정신장애진단도구(K-CIDI) 분류항목과 사용 질문 개수

<Table 1>은 데이터 전처리 과정을 나타낸다. 질문별 문장 단일화를 통해 데이터셋 입력 형식 통일하였다. 질문을 구성하는 문장이 2개 이상인 경우, 연구자가 질문을 읽고 원본 질문의 내용을 유지하는 한 개의 문장으로 수정한 과정을 나타낸다. 문맥 조건 추가를 통해 누락된 정보를 보충하였다. 이전의 질문과 이어져 반복되는 내용이 지시대명사로 대체된 경우, 후속 질문에 생략된 표현을 추가하였다. 전처리 과정에서 문장 본래의 의미나 목적이 달라지지 않았음을 확인하였다. 질문에 포함된 특수 기호 '(', '/', ')'는 삭제하였다.

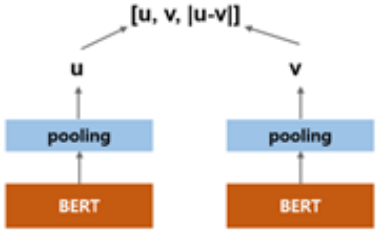
<Table 1> 질문별 문장 단일화와 문맥 조건 추가 예시

예시	설문문항	수정 전	수정 후
질문별 문장 단일화	E1	이제 슬프거나, 공허하거나, 우울하게 지냈던 때에 대해 질문하겠습니다. 지금까지 사는 동안, 2주 이상 거의 매일 하루 종일 슬프거나, 공허하거나, 우울하게 지낸 적이 있습니까?	지금까지 사는 동안, 2주 이상 거의 매일 하루 종일 슬프거나, 공허하거나, 우울하게 지낸 적이 있습니까?
	J17_11	술을 끊거나 줄이고 난 뒤, 처음 며칠 동안에 겪을 수 있는 몇 가지 문제에 대해 질문하겠습니다_경련성 발작을 했습니까?	술을 끊거나 줄이고 난 뒤, 처음 며칠 동안 경련성 발작을 했습니까?
질문별 문맥 조건 추가	E10_II	그런 기간들 중에, 2주 이상, 거의 매일, 말하거나 움직이는 게 평소보다 느려진 적이 있습니까?	우울했거나, 흥미를 잃었거나, 기운이 없었던 때, 2주 이상, 거의 매일, 말하거나 움직이는 게 평소보다 느려진 적이 있습니까?
	K24	그 일 이후, 계속해서 그 일에 대한 나쁜 꿈이나 악몽에 시달렸습니까?	스트레스나 충격적인 체험을 하고 난 이후, 계속해서 사건에 대한 나쁜 꿈이나 악몽에 시달렸습니까?

2.2. 문장 임베딩(sentence embedding)

트랜스포머[2] 모델 아키텍처 중 인코더 층(encoder layer)만으로 구현된 BERT (bidirectional encoder representations from transformers, [5])는 마스크 언어 모델링 (masked language modeling) 기법을 통해 문맥을 양방향(bidirectional)으로 학습한다. 선행문장과 후행문장을 판별하는 태스크와 마스킹된 단어를 예측하는 두 종류의 자기주도 지도학습(self-supervised learning)을 통해 문맥을 이해하는 능력을 습득한다.

BERT는 총 세 가지 방법으로 문장에 대한 임베딩 벡터를 구할 수 있다. 첫 번째로[CLS] 토큰의 임베딩 벡터를 활용하는 방법이 있다. [CLS] 토큰에는 문장을 구성하는 전체 토큰들의 종합적인 표현이 담기게 된다. 그러나 파인 튜닝(fine-tuning)을 하지 않은 모델은[CLS] 토큰이 특정 작업의 세부적인 의미를 정확하게 반영하지 않을 수 있다. 두 번째로, 모든 토큰의 임베딩 벡터를 평균 풀링(pooling)하는 방법이 있다. 이는 문장을 구성하는 모든 토큰에 대한 전체적인 문맥에 관심을 둔다. 세 번째로, 모든 토큰의 임베딩 벡터를 최대 풀링하는 방법이 있다. 이는 문장을 구성하는 모든 토큰 중 가장 핵심적인 문맥에 관심을 둔다. BERT는 cross-encoder 구조를 기반으로 두 개의 문장 쌍을 한 번에 모델에 입력 받아 문장 간 연관성을 계산하여 개별 문장의 임베딩을 계산하지 못한다는 한계가 있다. sentence-BERT[6]는 독립적인 문장 임베딩 벡터를 생성할 수 있도록 사전학습된 BERT를 파인 튜닝하였다.



<Figure 2> Sentence-BERT의 구조도

<Figure 2>는 sentence-BERT의 구조도를 나타낸다. BERT와 달리 sentence-BERT는 cross-encoder 구조를 통해 두 개의 문장을 각각 모델에 통과시켜 독립적인 문장 임베딩 벡터를 구한다. 이를 각각 u와 v라고 놓으면 u벡터와 v벡터의 차이의 절대값 벡터 |u-v|를 연결(concatenation)한다. 이후 문장 쌍 분류 또는 회귀 태스크를 통해 모델을 파인 튜닝한다. 문장 쌍 분류 태스크는 샴 네트워크 아키텍처를 활용해 두 문장의 유사도 확률을 소프트맥스 함수에 입력한 후 교차 엔트로피 손실(cross-entropy loss)을 최소화하는 방향으로 가중치를 업데이트한다. 문장 쌍 회귀 태스크는 트리플넷 네트워크 아키텍처를 활용해 기준(anchor) 문장과 긍정 문장(함의), 부정 문장(모순)의 세 가지 문장에 대해 기준 문장의 긍정 문장과의 유사도는 높아지는, 부정 문장과의 유사도는 낮아지는 방향으로 학습한다. 이러한 추가 학습을 거친 sentence-BERT는 문장의 문맥과 의미를 효과적으로 학습할 수 있다.

2.3. 사전 학습된 모델 선정

대규모의 텍스트 데이터에 대해 사전학습된 모델을 수행하고자 하는 특정 태스크에 전이학습(transfer learning) 또는 파인 튜닝하는 방식이 자연어처리 분야에서 우수한 성능을 보이고 있다. 대규모 데이터셋에 대한 모델 학습은 상당한 컴퓨팅 자원과 비용, 시간이 소요되는데,

Huggingface(https://huggingface.co/) 플랫폼에서는 자연어 이해(natural language understanding, NLU) 및 자연어 생성(natural language generation, NLG)을 위한 다양한 범용 아키텍처를 배포하고 있다. 따라서 사용자는 모델을 직접 학습시킬 필요없이 사전 학습된 모델의 가중치를 공유받아 임베딩을 추출하거나 다운스트림 태스크(downstream task)를 위한 추가 학습을 수행할 수 있다.

설문조사 데이터에 대한 문장 임베딩 벡터 생성에는 한국어를 포함하는 다국어 sentence-BERT 모델을 사용하였다. 파생 모델별로 학습 데이터셋의 종류 및 크기, 세부 아키텍처, 파라미터 개수 등 학습조건이 다르기 때문에 동일한 입력에 대해 서로 다른 임베딩을 출력한다. 본 분석에 적합한 모델을 선정하기 위하여, 입력에 대한 문장 임베딩 벡터 출력값을 네 가지 모델에 대해 비교하였다. 모델의 구체적인 구성(configuration)은 <Table 2>와 같다.

<Table 2> sentence-BERT 모델별 configuration

Pre-trained model	Base model	Dimension of embedding
paraphrase-multilingual-MiniLM-L12-v2	Bert	384
distiluse-base-multilingual-cased-v1	DistilBert	768
paraphrase-multilingual-mpnet-base-v2	XLMRoberta	
xlm-r-100langs-bert-base-nli-stsb-mean-tokens		

문항별 문맥 반영에 적합한 모델을 선택하기 위하여, 모델이 긍정의 의미를 표현하는 '예'와 부정의 의미를 표현하는 '아니오'의 의미를 이해하고 있는지와 두 단어의 차이를 임베딩으로 반영하는지에 대해 코사인 유사도를 통해 비교하였다. 비교 1에서는 랜덤하게 선택한 다섯 가지의 질문(항목별 1개)에 대해 응답 '예'와 '아니오'를 질문에 각각 결합한 두 문장을 생성하여 이들의 문장 임베딩 벡터 간 코사인 유사도를 산출하였다. 비교 2에서는 응답이 문장 임베딩에 미치는 영향을 비교하였다. <Table 3>은 비교 실험에 대한 결과를 나타내는데, 유사도가 낮을수록 '예'와 '아니오' 응답의 의미가 가장 크게 대비되는 모델을 의미한다. 모델 xlm-r-100langs-bert-base-nli-stsb-mean-tokens는 '예'와 '아니오'가 질문에 결합되었을 때, 다른 모델들에 비해 응답의 긍정과 부정 의미를 이해하고 있음을 확인할 수 있다. 비교 2의 실험결과로부터 B1A_A의 질문에 '예'와 '아니오'라고 응답한 경우는 0.770으로 다른 모델들보다 낮은 값을 가지고, 질문이 유사한 B1A_A와 B1A_E의 유사도는 '예'와 '아니오' 각각의 경우에 높은 유사도를 보이는 특징을 보인다. 이는 '예'와 '아니오'가 문장에서 가지는 의미가 모델을 통해 적절한 가중치를 학습하여 임베딩 벡터에 반영되어 있음으로 해석할 수 있다. 설문문항의 높은 신뢰도를 산출하기 지표인 크론바흐 α지수에서는 유사설문문항에 대한 높은 유사도를, 비유사문항에 대해서는 낮은 유사도를 가지는 특징을 가지는 설문문항을 선호하는데, 연구에서 제안하는 방법은 이러한 점과 맥락을 같이 한다고 해석할 수 있다.

이처럼 모델이 어떤 문맥에 더 큰 가중치를 두고 학습하였는지에 따라 동일한 단어, 문장에 부여되는 임베딩 값이 달라짐을 확인할 수 있다. 본 연구에서는 응답에 따라 달라지는 의미 차이에 주목하고자 하며, 추가 학습 없이 사전학습된 모델을 활용하기 위하여 유사도 방향이 단어의 뜻에 의해 좌우되는 xlm-r-100langs-bert-base-nli-stsb-mean-tokens 모델을 활용한다.

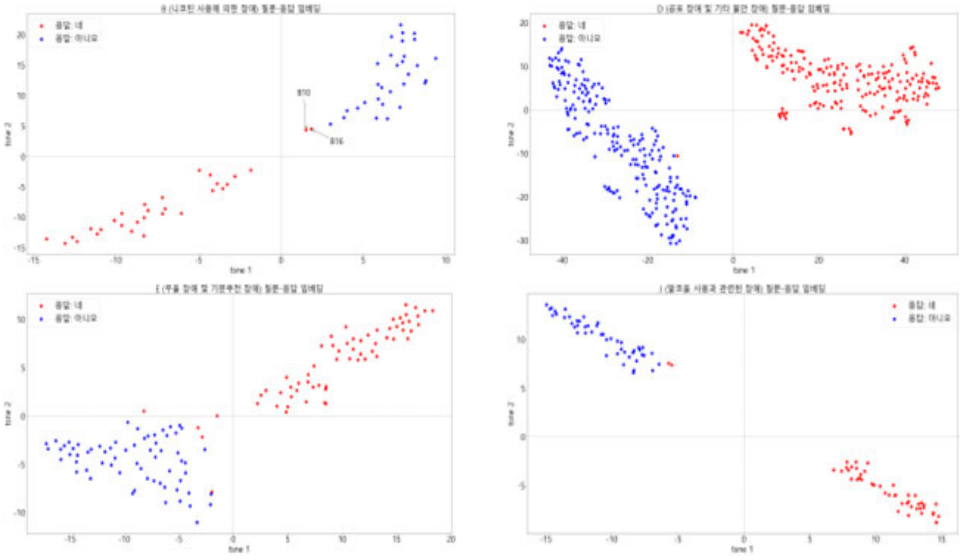
<Table 3> 세 가지 비교 셋팅에 따른 문장 임베딩 코사인 유사도

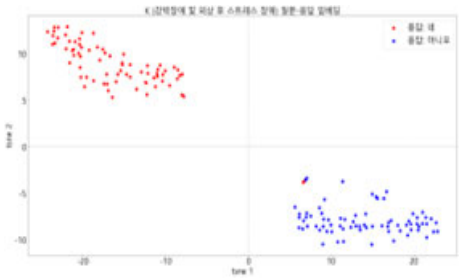
				모델명			
유사도 방향성		조건		paraphrase-multilingual-MiniLM-L12-v2	distiluse-base-multilingual-cased-v1	paraphrase-multilingual-mpnet-base-v2	xlm-r-100langs-bert-base-nli-stsb-mean-tokens
비교 1	↓	질문-[예] / 질문-[아니오]	B14	0.982	0.990	0.996	0.730
			D65_2	0.989	0.991	0.995	0.565
			E2	0.979	0.994	0.995	0.582
			J7B	0.980	0.981	0.995	0.801
			K22_1_I	0.993	0.991	0.997	0.539
비교 2	↓	B1A_A-[예]/ B1A_A-[아니오]		0.976	0.991	0.995	0.770
	↑	B1A_A-[예] /B1E_A-[예]		0.930	0.696	0.958	0.977
	↑	B1A_A-[아니오]/B1E_A-[아니오]		0.933	0.683	0.958	0.998

3. 분석결과

3.1. 질문-응답 쌍 그룹화

응답에 따른 질문의 임베딩 공간을 해석하기 위하여 <Table 4>를 근거로 모델 xlm-r-100langs-bert-base-nli-stsb-mean-tokens를 사용한다. 전처리를 마친 질문에 응답 '네'와 '아니오'를 매핑하여 하나의 질문에 대해 두 개의 입력 데이터 쌍을 생성하였다. 각 설문 문장 임베딩 벡터는 평균 풀링으로 구한 768차원의 벡터를 산출하였으며, t-SNE[3]를 활용해 2차원으로 축소하여 항목별로 시각화하였다. 5가지 유형에 대한 질문-응답 임베딩의 t-SNE 차원축소 시각화는 <Figure 3>에 나타낸다. 시각화 결과, 응답 '네'와 '아니오'가 결합된 질문은 t-SNE 2차원 차원 축소 임베딩 공간상에서 대부분 뚜렷한 구분을 보인다. 이는 문장 임베딩 벡터가 반대되는 응답의 의미를 반영하고 있음을 알 수 있다.





<Figure 3> 유형별 질문-응답 임베딩t-SNE 차원 축소 시각화

<Table 4> 반대 응답과 가까운 거리의 설문문항과 해당 질문

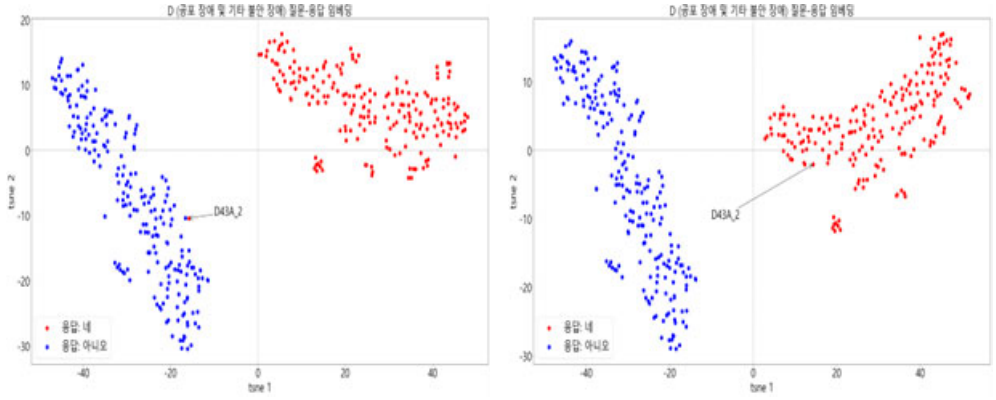
설문문항	질문
B10	담배를 끊거나 줄인 뒤 생긴 문제 때문에, 일을 하거나 다른 사람들과 어울리는 데 지장을 받은 적이 있습니까?
B16	담배를 피우기 위해서, 운동이나 일, 친구나 친척들과의 모임 같은 중요한 활동을 포기하거나 많이 줄인 적이 있습니까?
D43A_2	버스, 기차, 또는 자동차를 타고 여행하는 상황을 심하게 두려워 한 적이 있습니까?
E8_II	(우울했거나/흥미를 잃었거나/기운이 없었던) 때, 2주 이상, 거의 매일 밤, 잠이 안 오거나, 한밤중에 깨거나, 너무 일찍 깨 잠자기가 어려웠습니까?
E11_II	(우울했거나/흥미를 잃었거나/기운이 없었던) 때, 2주 이상, 가만히 앉아 있지 못하고 왔다 갔다 하거나, 앉아 있어도 손을 가만히 두지 못하는 등 안절부절 못한 적이 있습니까?
E15A_II	(우울했거나/ 흥미를 잃었거나/기운이 없었던) 때, 집중할 수가 없어서, 평소 좋아하던 책, TV, 또는 영화를 볼 수 없었습니까?
E19_II	(우울했거나/ 흥미를 잃었거나/기운이 없었던) 때, 너무 우울해서, 자살해야겠다는 생각을 많이 했습니까?
E22_II	(우울하거나/흥미를 잃거나/기운이 없이) 지냈던2주 동안, 거의 매일, 아침에 일어날 때는 특히 기분이 나쁘다가 시간이 지날수록 기분이 나아졌습니까?
E26A	(우울하거나/흥미를 잃거나/기운이 없이) 지냈던 기간들 중에, 2주 이상 우울함과 관련된 증상 때문에 직장생활을 하거나, 집안, 가족, 또는 자신을 돌보는 데 지장이 많이 있었습니까?
J6	지금까지 사는 동안, 술에 취하거나 술에서 덜 깨서 학교, 직장, 또는 가정 생활에 지장을 자주 받은 적이 있습니까?
J16	술 때문에 운동, 학업, 직업, 또는 친구나 친척과의 관계유지를 포기하거나 많이 줄이는 등 중요한 활동을 포기하거나 많이 줄인 적이 있습니까?
K7	원하지 않는 불쾌한 생각 또는 떨쳐버릴 수 없는 불쾌한 생각 때문에 생활이나 일에 지장을 받거나, 친척이나 친구들과 지내는 데 어려움이 생기거나, 또는 기분이 몹시 상했습니까?
K19	어리석은 행동을 반복하거나 일정한 순서대로 행동하거나 개수를 세는 행동 때문에 생활이나 일에 지장을 받거나, 친척이나 친구들과 지내는 데 어려움이 생기거나, 또는 기분이 몹시 상한 적이 있습니까?

<Table 4>는 반대 응답과 가까운 거리의 문항을 나타낸다. 이들 변수에 등장하는 친구, 친척, 여행, 가족, 생활, 아침 등의 단어의 임베딩이 편향되었을 가능성을 발견하였다. 언어모델은 확률적으로 공동 등장(co-occurrence) 정보를 학습하기 때문에 해당 단어들과 함께 등장한 주변 맥락에 의해 영향을 받는다. 즉, 긍정적인 감정이나 경험과 연관되어진 임베딩 공간을 학습하는 경우 단어의 중립성이 훼손되어 설문문항이 묻고자 한 본래의 의도가 왜곡될 수 있다. 이의 근거로, 보다 중립적인 의미의 단어로 수정하거나 일부 단어를 삭제한 후 임베딩 공간을 비교하였다. <Table 5>는 일부의 예시로 'D43A_2' 문항과 'J6' 문항을 원본질문의 표현을 수정하여 문장의 의미가 긍정과 부정의 의미의 방향을 가지게 하였다. <Figure 4>는 임베딩으로 'D43A_2'의 문항의 위치가 수정된 것을 확인할 수 있다. 문장의 의미를 유지하며 단어들을 변

경하였을 때 문장의 임베딩 공간이 달라짐을 확인할 수 있다. 단, 연구자가 임의로 단어를 변경하는 과정에서 주관적인 판단으로 인해 성능이 저하될 가능성이 존재한다. 단어의 왜곡이 발견될 경우, 설문의 목적과 문항의 의도를 유지하면서 연구자의 주관적인 요소를 배제하여 단어 선택의 객관성을 확보할 필요가 있다.

<Table 5> 수정 전후 질문

설문문항	원본 질문	변경 질문
D43A_2	버스, 기차, 또는 자동차를 타고 여행하는 상황을 심하게 두려워 한 적이 있습니까?	버스, 기차, 또는 자동차를 타고 이동하는 상황을 심하게 두려워 한 적이 있습니까?
J6	지금까지 사는 동안, 술에 취하거나 술에서 덜 깨서 학교, 직장, 또는 가정 생활에 지장을 자주 받은 적이 있습니까?	지금까지 사는 동안, 술에 취하거나 술에서 덜 깨서 생활에 지장을 자주 받은 적이 있습니까?

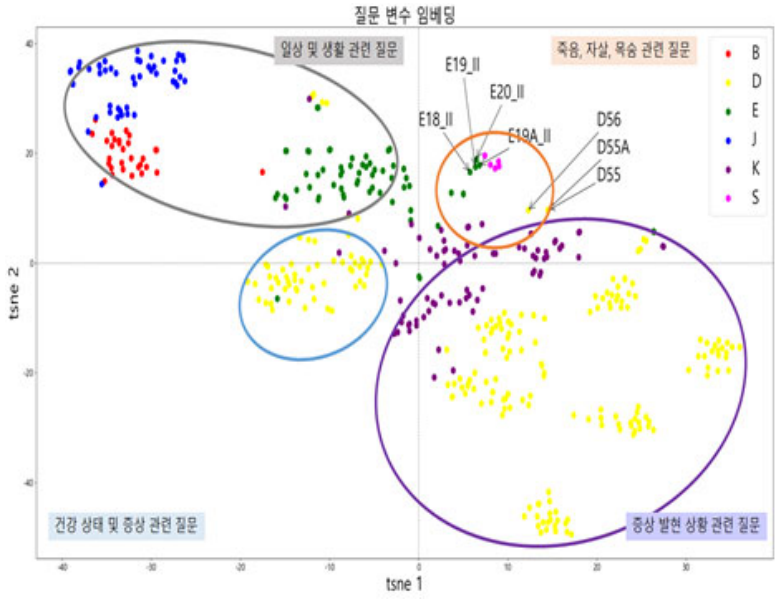


<Figure 4> D(공포 장애 및 기타 불안 장애) 수정전후 질문-응답 임베딩t-SNE 차원 축소 시각화

3.2. 변수그룹별 질문

응답을 기준으로 그룹화와 동일한 방법으로 설문문항에 따른 질문의 임베딩 공간을 해석해 보기로 한다. <Figure 5>는 전처리를 마친 질문을 입력 데이터로 하여 768차원의 문장 임베딩 벡터는 t-SNE를 활용해 2차원으로 축소하여 시각화를 나타낸다.

제1사분면에 위치한 S(자살) 항목의 질문 변수들은 일부 E(우울장애 및 기분부전 장애) 항목의 질문 변수들과 근접한데, 이에 해당하는 질문 변수 E18_||, E19_||, E19A_11, E20_||에는 공통적으로 '자살'과 '죽음' 단어가 등장하였다. 또한 이들과 밀접한 거리에 이는 D(공포 장애 및 기타 불안 장애)에는 '목숨이 위험'이라는 어구가 확인되었다. 이를 통해 1사분면 공간에는 죽음, 자살, 목숨과 직접적으로 관련 있는 변수들의 문맥이 포함되어 있음을 유추할 수 있다. B, D, E, J, K, S의 분류항목의 설문문항들이 그룹을 형성하고 있음을 확인할 수 있는데, 특히, B의 니코틴 사용장애와 J의 알코올 사용장애가 다른 그룹들보다 상대적으로 가깝게 분포하고 있다. D의 공포 장애 및 기타 불안 장애는 비교적 넓은 범위에 분포하고 있음을 알 수 있다. 가장 밀집하게 분포하고 있는 문항들은 S(자살) 설문문항들로 이들과 인접한 거리에 있는 다른 영역의 설문문항들은 <Table 6>에 정리하였다. D(공포 장애 및 기타불안 장애), E(우울장애 및 기분부전 장애)에 분류된 문항들로 이들 문항의 기술에는 직접적으로 '자살', '죽음', '목숨'과 같은 단어들이 사용되었음을 알 수 있다.



<Figure 5> 질문 변수 임베딩 t-SNE 차원 축소 해석

<Table 6> S(자살) 설문문항군과 근접한 임베딩 공간의 질문

설문문항	질문
D55	목숨이 위험한 상황에서 불안발작이 있습니까?
D55A	목숨이 위험하지 않은 상황에서도 불안발작이 있습니까?
D56	목숨이 위험하지 않은 상황에서 전혀 예상하지 못한 불안발작을 두 번 이상 겪은 적이 있습니까?
E18_II	우울했거나, 흥미를 잃었거나, 기운이 없었던 때, 2주 이상, 죽음에 대한 생각을 많이 한 적이 있습니까?
E19_II	우울했거나, 흥미를 잃었거나, 기운이 없었던 때, 너무 우울해서, 자살해야겠다는 생각을 많이 했습니까?
E19A_II	우울했거나, 흥미를 잃었거나, 기운이 없었던 때, 자살을 어떻게 할 지 계획을 세웠습니까?
E20_II	우울했거나, 흥미를 잃었거나, 기운이 없었던 때, 죽음에 이를 가능성이 없어 보이는 시도도 모두 포함하여 자살을 시도 했습니까?

제2사분면에 위치한 B(니코틴 사용에 의한 장애)와 J(알코올 사용과 관련된 장애) 항목은 매우 근접한 공간에 위치하고 있다. 이는 문장에 반복적으로 등장하는 '담배'와 '술'의 임베딩 공간이 유사하기 때문인 것으로 확인된다. 특히, B, J, E에는 일상과 관련된 생활 패턴 및 평상적 상황에 대한 질문이 다수 존재하였다. 이를 통해 제2사분면 공간에는 일상 및 생활과 관련 있는 변수들의 문맥이 포함되어 있음을 유추할 수 있다. 제3사분면에 위치한 D(공포 장애 및 기타 불안 장애) 변수들은 건강 상태 및 증상과 관련된 질문들이 군집되어 있다. 걱정 또는 긴장한 상태에서 발현된 구체적인 증상에 대한 질문이 다수 존재하였다. 특히, K27의 '땀이 나거나 심장이 빨리 뛰거나 몸이 떨리다'와 같은 특정한 신체적 현상에 대한 질문이 이와 유사한 거리에 위치하고 있음을 알 수 있다. 이를 통해 3사분면 공간에는 건강 상태 및 증상과 관련 있는 변수들의 문맥이 포함되어 있음을 유추할 수 있다. 제4사분면에 위치한 D 항목 변수와 K(강박장애 및 외상 후 스트레스 장애) 항목 변수들은 공포심 또는 불안 증상이 발생하는 특수한 상황과 관련된 질문들이 군집되어 있다. 특정 증상의 발현을 유발하는 구체적인 상황에 대해 묘사하고 있으며 경험한 강박 증상에 관련된 특수한 상황에 대해 한 질문이 다수 존재하였다. 이를 통해 4사분면 공간에는 증상 발현 상황과 관련 있는 변수들의 문맥이 포함되어 있음을 유추할 수 있다.

3.3. 문장 임베딩을 활용한 정신건강 집단 분류

질문-응답 임베딩 벡터가 정신건강 집단 타겟 변수 분류에 미치는 영향력을 살펴보기 위해 실제 설문조사 데이터를 활용하였다. 먼저, 변수명과 숫자로 범주화되어 있는 응답(1: 아니오, 5: 네)은 각각 자연어 형태로 변환하였다. S1(자살하는 것에 대해 진지하게 생각한 적이 한번이라도 있습니까?) 변수의 응답을 기준으로 정신건강 정상군과 위험군을 라벨링하고, S1 변수에 '네'라고 응답한 위험군(class: 1) 그룹과 '아니오'라고 응답한 정상군(class: 0) 그룹의 이진 분류분석(binary classification)을 수행하였다. 전체 응답자 5,511명 중 574명이 위험군에 포함되며, 불균형한(imbalanced) 데이터 분포를 고려하여 정상군 574명을 10회 랜덤 샘플링하였다. <Table 7>의 3가지 형태로 데이터셋을 구성하였으며 샘플링된 정상군과 위험군 10개 집단에 대해 XGBoost 모델[7]을 적용하여 반복 실험하였으며 분류정확도로 비교하였다.

<Table7> 정신건강 집단군 분류 실험을 위한 데이터셋 구성

XGBoost classification		
Without context		With context
원시 설문조사자료 427개 변수	통계적 추정 유병률 생성변수 60개 변수	임베딩 벡터 변수 427개 변수

문장 임베딩 벡터 데이터셋은 그룹 별 574명, 총 1,148명에 대해 각 427개의 질문-응답 쌍을 매핑(S, 자살 유형 3개 질문 포함)하였으며, 5개의 분류항목별(B, D, E, J, K)로 t-SNE 차원 축소를 수행하였다. 총 1,148명에 대해 t-SNE 1차원 데이터셋은 427차원의 학습 데이터를, t-SNE 2차원은 한 설문문항 당 2개의 차원을 가지므로 총 854차원을 가진 학습 데이터를 생성하였다. 일반적으로 언어모델에서는 perplexity이 큰 값으로 설정할수록 훈련시 모델의 언어 이해력(예측확률)이 높은 모델을 선호한다는 의미를 가지는데, t-SNE 학습시에는 일반적인 거리계산시 인접한 데이터에 대한 상대적인 거리를 조절할 수 있는 인자로 활용되어 국소성(locality) 정도를 조율하는 모수로 활용된다. 이 옵션이 낮을수록 거리계산시 국소성이 더 강화된다고 할 수 있다. 연구에서는 두 값을 설정하여 t-SNE 축소 좌표를 산출하는데 반영하였다.

생성변수 데이터셋은 동일한 그룹에 대해 1년 유병률, 1개월 유병률, 평생 유병률 변수를 사용하였다. 단, 자살 관련 행동에 해당하는 항목 변수 중 SIdea는 타겟정보를 포함하고 있어서 학습 데이터에서 제외하였다. 또한, SIdea 이외의 자살 관련 행동 변수 8개와, 문장 임베딩 항목에 반영되지 않은 인터넷 게임장애 항목 9개와 게임 생활 습관 항목 1개도 제외하였다. 따라서 1,148명에 대해 60개의 생성변수를 학습 데이터로 설정하였으며, y는 동일하게 그룹별로 라벨링하였다. 각 실험마다 훈련 데이터와 테스트 데이터의 비율(split_random_state)은 8:2로 설정하였으며, min-max 정규화를 적용하였다. 모델의 파라미터는 각각 트리 모델의 개수(n_estimators) 1000, 트리의 최대 깊이(max_depth) 3으로 설정하였다.

<Table 8> XGBoost 정확도(accuracy) 결과(10회 반복의 평균과 표준편차)

	설문문항	생성변수	t-SNE 2차원
평균(표준편차)	0.705(0.021)	0.674(0.016)	0.827(0.021)

<Table 8>은 10번의 반복실험에 대한 결과를 나타낸다. 생성변수로 학습한 모델은 평균 0.674의 정확도를 보였으며, 설문문항의 응답에 대한 정확도는 0.705이다. 질문-응답 쌍의 문장 임베딩 벡터를 반영한 모델은 t-SNE 2차원 축소 데이터에서 분류 성능이 0.827로 평균적으

로 가장 높은 것으로 나타났다. 설문문항의 응답과 설문문항의 문장의 중요도를 반영할 경우를 비교할 때, 최대 0.153의 정확도가 상승할 수 있음을 알 수 있다. <Table 9>는 문장 임베딩 벡터 데이터셋으로부터 구한 XGBoost 모델의 피처 중요도를 살펴보기 위한 SHAP(shapley additive explanations)[8] 값으로, 고위험군과 저위험군 두 클래스 분류 모델의 예측값에 미치는 절대 기여도가 높은 상위10개의 변수들을 확인할 수 있다. K 항목의 특정 사건 또는 스트레스로 인한 증상, D 항목의 불안 관련 증상이 클래스 분류에 주요한 기여를 한 것으로 생각된다. 또한, 저위험군에서도 빈번히 발견되는 B 항목과 J 항목에 대해, B7_2, B6, J18_1처럼 특정 증상이 발견될 경우에는 중독성 물질 의존도가 높은 상태이며, 정신적 문제의 연관성이 높을 가능성이 있다고 해석할 수 있다. E 항목의 변수가 모든 실험에서 높은 기여도를 보이는 것으로 보아 우울 관련 증상이 다른 항목들에 비해 정신 건강에 미치는 영향이 상대적으로 높다고 볼 수 있다.

<Table 9> t-SNE 2차원 벡터 활용시 10회 반복에서의 예측기여도 상위 설문문항

설문문항	질문
E3_II	2주 이상(슬프거나, 공허하거나, 우울하게 지냈던/ 또는 흥미를 잃고 지냈던) 때에, 아주 힘들게 일한 것도 아닌데, 거의 매일, 항상 기운이 없거나 피곤했습니까?
E2	사건으로 인해 생긴 문제 때문에, 잔치, 사회적 행사나 모임에 참석하지 못한 적이 있습니까?
B7_2	지금까지 사는 동안, 2주 이상 일, 취미, 또는 평소 좋아하던 것들 대부분에 흥미를 잃은 적이 있습니까?

참고로, 부록의 <Table 10>은 정신장애진단도구(K-CIDI) 설문문항 중 분석에 사용된 설문문항 정보를 나타낸다. 집단분류시에 집단의 예측에 직접적으로 기여할 수 있는 설문문항은 최대한 배제하였으나 내용에 따라 일부 포함되었을 가능성은 있을 수 있음을 미리 밝힌다.

4. 결론 및 제언

본 분석에서는 보건복지부 국립정신건강센터에서 수집하는 정신건강실태조사 데이터를 대상으로 자살사고, 자살계획, 자살시도 대상자의 집단적 특성 파악을 목표로 하였다. 개별문항의 집계점수가 가진 정보의 축약성을 극복하고자, 설문문항의 문맥으로부터 위험의 크기를 변수화하는 방법을 제안 및 활용하였다. 일반적인 조사 방법론에서는 설문 문항들의 상관관계를 통해 잠재집단을 추출하여 잠재변수(factor)를 정의하는 과정에서 연구자의 주관적 해석이 개입되어야 한다. 본 연구에서 활용한 자연어 임베딩 방법론은 텍스트 데이터를 확률적으로 학습하여 문항 간의 관계를 정량적으로 추출할 수 있다. 이로부터 산출된 점수를 활용하여 XGBoost의 정상군과 자살위험군간의 분류 정확도를 통상적인 방법에 비해 개선할 수 있음을 확인하였는데, 단순 설문문항에 기초한 방법보다 최대20% 정도의 정확도 개선을 달성하였다. 특히, 상태나 동작이 이루어지는 강도를 나타내는 정도부사를 활용해 이분변수 이상의 척도로 적용 범위를 확장할 수 있다. 예를 들어5점 척도의 보기를 매우 그렇다/그렇다/보통이다/아니다/매우 아니다로 구성할 경우'매우'라는 표현의 차이가 임베딩에 반영되어 선택지의 구성 의도를 구체화한다. 이처럼 본 연구에서 제안하는 임베딩 기법은 단어와 문장의 의미를 다차원 벡터 공간에 매핑하여, 미묘한 뉘앙스와 문맥적 차이를 반영할 수 있다.또한, 질문별 자살과의 상관관계를 가중치로 하여 문장 임베딩 벡터를 학습시킨다면 고위험군 식별에 도움이 되는 질문 변수를 추출하는 데 용이할 것으로 판단된다. 자연어의 문장 임베딩에는 질문과 응답의 문맥적 관계가 반영되기 때문에, 질문의 깊이에 따라 다른 임베딩 값을 부여할 수 있고 위험군 판별에 주요한 질

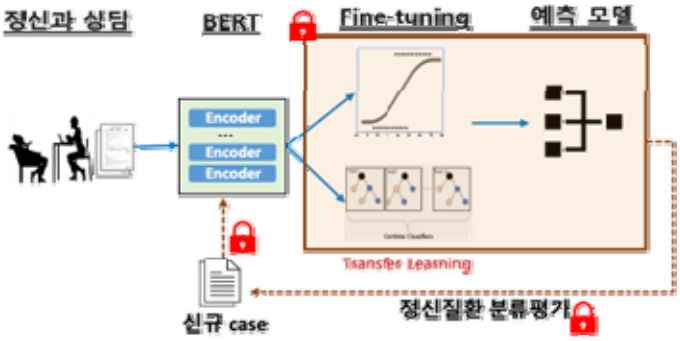
문에는 더 큰 가중치를 두고 학습할 수 있을 것으로 기대된다. 따라서 질문으로부터 획득할 수 있는 정보의 양이 증가할 수 있으며, 추가적으로 자연어의 문맥적 의미를 반영할 수 있다. 이러한 분석 결과를 바탕으로, 정신건강 관련 텍스트 데이터로 사전 학습된 모델을 파인 튜닝(fine-tuning)하거나 프롬프트 튜닝(prompt-tuning)을 적용한다면 더욱 정확한 문장 임베딩 벡터를 추출할 수 있을 것으로 기대된다. 더욱이 충분히 학습된 모델을 보유한다면, 설문문항으로부터 정신건강 질환 항목의 심각성의 정도 또는 질환 보유 현황을 측정하는 데 미치는 영향의 크기를 정량적으로 해석할 수 있을 것이라 기대한다. 다만, 본 연구에서는 정상군과 자살위험군 간의 인구통계학적인 변수정보를 모델에 반영하지 못하였는데 이를 반영하면, 결과의 정확도를 개선할 수 있을 것으로 기대한다.

향후, 정신건강 데이터분석에 기초하여 연구방향을 다음으로 정리할 수 있다. 잠재변수 추출 기법과 본 연구의 자연어 임베딩 방법론을 함께 활용하여 각 기법의 성능과 유용성을 비교 분석하는 후속 연구를 발전 방안으로 제안하여 설문조사의 설계와 분석에서의 관계를 보완할 수 있을 것으로 기대한다. 또한, 설문 문항의 워딩이 응답의 질에 미치는 영향을 정량적으로 평가함으로써, 질 높은 응답을 유도할 수 있는 최적의 질문 구성 방식을 탐구하는 방향으로 연구를 확장할 수 있다. 특정 단어 조합이나 문장의 구조가 응답자의 심리적 부담을 줄이거나, 더 구체적이고 신뢰도 높은 정보를 끌어낼 수 있는지 분석하여 방법론의 근거를 더욱 강화할 수 있을 것이다.

최근 이상동기 범죄(묻지마 범죄)나 무차별 범죄 등이 급증하고 있는데 외래치료 명령제, 외래치료 지원제도로 자/타해 위험성이 있는 개인을 관리하는 제도의 내실있는 도입이 필요할 것으로 사료된다. 기존에 운영 중인 여러 프로그램들에서 획득되는 정형, 비정형의 다양한 기록데이터는, 예를 들어, 주소, 이름, 채팅 메시지, 이메일 주소, 사건기록, 약물처방기록, 학적부 생활기록, 개인 신병 서술 등의 상담기록(텍스트), 극 민감정보를 직접적으로 담고 있는 경우가 많으므로 프라이버시 이슈가 지속적으로 발생할 수 있다. 개인정보를 보호할 수 있는 예측모델링 개발이나 방안이 필요하다고 할 수 있다.

본 연구에서는 사전학습된 모델을 정신건강자료를 이해하는 언어모델로 임베딩을 추출해 특징을 추출하여 변수화하는 방안이 유효한 방법으로 활용될 수 있음을 확인하였는데, 예측모델링을 고도화하기 위해서 극 민감정보를 활용하는 추가적인 방안이 필요하다. <Figure 6>은 주어진 텍스트정보로부터 sentenceBERT로 임베딩하고 이를 암호화(encryption)하여, 동형암호화(homomophic encryption)된 상태에서 로지스틱 회귀분석을 적용하는 모식도를 나타낸다. 최종 스코어값은 암호화되어 있으므로 암호키로 복호화(decryption)하면 실제값을 확인할 수 있다. 텍스트 원본은 노출없이 로지스틱 회귀분석을 통한 위험도를 산출할 수 있음을 의미한다[9,10]. 정신질환자 질적상담자 인터뷰 데이터를 BERT로 파인-튜닝하여 정신건강자료를 이해하는 언어모델을 구축하고, 모델을 암호화한 후, 웹 등으로 인입(평가요청)되는 암호화된 [CLS]벡터와의 연관성을 추론하여 정신질환자 텍스트 유사도추출(정신질환기능평가) 평가할 수 있다. 서비스이용자는 원문을 공개하지 않고도 정신건강평가 모델을 통해 유사도 측정할 수 있으며, 기존 대표텍스트(barycenter)와의 유사도평가도 가능하다. 또한, 평문에서 정신건강 관련된 서비스기능별로 모델을 구축할 경우 다차원의 유사도 측정이 가능하다. 정신건강실태조사에서의 경우에서는 B/C/D/E와 같은 영역이 이에 해당된다. 한편, 데이터연계나 결합을 통해 여러 기관으로부터 원본데이터가 아닌 임베딩 벡터를 제공받고 이 정보를 앙상블하여 의사결정(decision model)을 구축할 수 있다. 우울증 등의 음성기록이나 영상 등의 자료는 동일한 방식으로 각각 wav2vec[11]과 ViT[12]와 같은 트랜스포머를 활용한 특징추출기와 같은 다양한 backend 모델을 통해 전이학습이 가능하다. 이는 다기관 데이터를 멀티모달리티

(multimodality) 관점에서 해석할 수 있음을 시사하며, 이러한 연구는 향후 연구주제로 남기고자 한다.



<Figure 6> 동형암호화(homomorphic encryption)된 상태에서 로지스틱 회귀분석을 적용하는 모식도

사사(Acknowledgements)

본 연구는 보건복지부 2021년 정신건강실태조사 마이크로데이터(NMHSK-21)를 활용한 것으로, 연구의 결과는 보건복지부와 관련없습니다.

본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(NO.2022M3J6A1084845)이며, 서울시립대학교 도시과학빅데이터AI연구원의 슈퍼컴퓨팅 자원을 지원 받아 수행되었습니다.

This study utilized data of the National Mental Health Survey 2021(NMHSK-21). The results of this study is irrelevant with the Ministry of Health and Welfare of South Korea.

This research was supported by the National Research Foundation of Korea(NRF) grant funded by the korea government(MSIT) (NO.2022M3J6A1084845). The authors acknowledge the Urban Big Data and AI Institute of the University of Seoul supercomputing resources (http://ubai.uos.ac.kr) made available for conducting the research reported in this paper.

참고문헌(References)

[1] Kim DB, Ahn IK. 2004. A Study on the Concept of Mental Health in Korea. Korean Journal of Social Welfare 56(1), 203-233.

[2] Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention is all you need. Advances in Neural Information Processing Systems 30. pp. 5998-6008.

[3] Maaten L, Hinton G. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9. pp. 2579-2605.

[4] Roweis S, Hinton G. 2002. Stochastic neighbor embedding. Advances in Neural Information Processing Systems.

[5] Devlin J, Chang M, Lee K, et al. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

[6] Reimers N, Gurevych I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Conference on Empirical Methods in Natural Language Processing.

[7] Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), Association for Computing Machinery, New York, pp, 785-794.

[8] Lundberg SM, Lee SI. 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30.

[9] Lee G, Kim M, Park J, et al. 2022. Privacy-preserving text classification on BERT embeddings with homomorphic encryption. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3169-3175.

[10] Lee S, Lee G, Kim J. 2023. HETAL: efficient privacy-preserving transfer learning with homomorphic encryption. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023) 202, pp, 19010-19035.

[11] Baevski A, Zhou H, Mohamed A, et al. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20) 1044, pp. 12449-12460.

[12] Dosovitskiy A, Beyer L, Kolesnikov A, et al. 2021. An image is worth 16x16 words: transformers for image recognition at scale. ICLR.