문서 임베딩 생성 방법에 따른 토픽 모델링 성능 비교 :S-BERT와 LLM 임베딩 모델 중심으로

(Comparison of Topic Modeling Performance by Document Embedding Generation Methods : Focused on S-BERT and LLM Embedding Models)

박지원* 양치복* 이충권*

(Ji Won Park, Chi Bok Yang, Choong Kwon Lee)

요 약

본 연구는 임베딩 생성 방법에 따른 토픽 모델링의 성능 비교를 위해 대규모 언어 모델(LLM) 임베딩과 S-BERT 임베딩을 활용하여 BERTopic 기반 토픽 모델링 실험을 수행하였다. BBC News, 20 Newsgroups, IMDB 데이터를 사용하여 텍스트를 전처리하고, CLS, mean, max 등을 조합한 7가지 풀링 방법에 따라 토픽 일관성(Cv, NPMI)과 토픽 다양성(topic diversity)을 평가하였다. 실험 결과, 성능 우위는 데이터셋과 임베딩 모델에 따라 달랐다. mean 풀링 기준으로는 LLaMA 임베딩이 상대적으로 안정적인 토픽 일관성을 보였으며, CLS·max를 포함한 듀얼 풀링을 적용하면 S-BERT 임베딩이 LLaMA 임베딩보다 Cv와 NPMI에서 높게 나타났다. 듀얼 풀링은 토픽 다양성을 저해하지 않으면서 전반적인 토픽 일관성 향상에 기여하였다. 이는 S-BERT 임베딩으로 적절한 문서 임베딩을 방법을 적용할 경우, LLM기반 임베딩과 비교하여 토픽 일관성 측면에서 우수하거나 대등한 성능을 달성할 수 있음을 시사한다. 또한 듀얼 풀링은 추가 연산이 단순 집계 수준에 그쳐 단일 풀링 대비 평균 소요 시간 차이가 거의 발생하지 않았다. 본 연구는 토픽 모델링 성능 개선을 위하여 고비용의 LLM기반 임베딩을 활용하는 것보다 임베딩 특성과 풀링 전략을 고려한 설계가 실질적인 대안이 될 수 있음을 보여준다.

■ 중심어 : 임베딩 ; 토픽 모델링 ; 거대 언어 모델 ; 듀얼 풀링

Abstract

This study compares topic modeling performance across embedding generation methods by applying BERTopic with S-BERT and large language model (LLM) embeddings on the BBC News, 20 Newsgroups, and IMDB datasets. We evaluate topic coherence (Cv, NPMI) and topic diversity under seven pooling strategies (including CLS, mean, and max), and find that while LLaMA embeddings yield relatively stable coherence under mean pooling, dual pooling strategies that include CLS and max often allow S-BERT embeddings to achieve equal or higher Cv and NPMI than LLaMA without substantially harming topic diversity. Because dual pooling requires only simple aggregation and adds little runtime overhead, these results suggest that, with appropriate pooling design, lower-cost S-BERT embeddings can in some cases replace or rival LLM-based embeddings for topic modeling.

■ keywords : Embedding ; Topic Modeling ; Large Language Model ; Dual Pooling

I. 서 론

인공지능(Artificial Intelligence, AI)의 고도화로 업무 보조·자동화의 활용이 확대되면서 국가와 산업 전반의 관심이 집중되고 있다. 그중에서도 텍스트와 이미지를 이해하는 자연어 처리

* 정회원, 계명대학교 경영정보학과

접수일자 : 2025년 11월 06일
수정일자 : 2025년 12월 13일
게재확정일 : 2025년 12월 21일
교신저자 : 이충권 e-mail : cklee@kmu.ac.kr

(Natural Language Processing, NLP) 기술이 빠르게 발전하며, AI가 사람의 언어와 시각적 정보를 효과적으로 해석할 수 있는 기반을 마련하고 있다[1,2]. 이러한 기술의 핵심은 비정형 정보를 수치화하여 인공지능이 이해할 수 있는 구조로 표현하는 임베딩(Embedding) 과정이다. 임베딩 품질이 높을수록 텍스트의 의미적 관계를 정확히 파악하고, 이미지의 시각적 특징을 정밀하게 인식할 수 있다[3].

다양한 산업에서는 도메인 전문용어나 문서를 사전학습 언어 모델(Pre-trained Language Model, PLM)을 활용하여 문서 임베딩을 생성함으로써 임베딩 품질을 높이고자 하였다[4, 5]. 그러나 기존의 임베딩 모델은 주로 일반 도메인 데이터를 기반으로 학습되어, 산업별 전문 용어나 기술 문서에서의 의미적 맥락을 충분히 반영하지 못하는 한계가 있다. 이러한 한계를 극복하기 위한 대안으로, 거대 언어 모델(Large Language Model, LLM)이 주목받고 있다. GPT나 Gemma 같은 거대 언어 모델(Large Language Model, LLM)이 Sentence-BERT(S-BERT) 같은 전통 임베딩보다 폭넓은 언어 패턴과 의미적 뉘앙스를 포착할 수 있어 번역, 문서 요약, 질의응답 등 다양한 영역에서 높은 성능을 보였다[6,7]. 특히, 최근 연구에서 LLM 임베딩을 적용한 토픽 모델링 연구들이 진행되었으며 [8-10], S-BERT나 DistilBERT와 같은 사전학습 언어 모델 기반의 임베딩보다 LLM 기반 임베딩을 활용하였을 때, 토픽 일관성이 향상되는 것으로 보고되었다.

하지만, 대부분의 자연어 처리 연구에서는 문서 임베딩을 생성할 때, 주로 평균 풀링(mean pooling) 방식을 사용하여 다양한 임베딩 풀링 기법에 대한 실증적 검증은 아직 부족하다[9]. 풀링은 사전학습 언어 모델이 토큰별 은닉 표현을 출력할 때 문서 수준의 고정 길이 벡터를 얻기 위한 필수적 단계이다. 풀링에는 CLS(Classification), 평균(MEAN), 차원별 최댓값(MAX)이 있으며, 각 방식에는 CLS의 대표성 저하, MEAN의 주제 신호 희석, MAX의 문맥 정보손실 등의 한계가 있다[11,12]. 최근 임베딩 풀링에 관련된 연구에서는 mean, max 풀링을 결합하여 표현 차원을 확장함으로써 특정 주제 신호 희석과 맥락 정보 손실에 대한 문장 정보손실을 완화하고자 하였다[12]. 그러나, 이러한 시도들은 주로 BERT 계열 인코더 모델을 대상으로 단일 데이터셋 수준에서 검증하였지만, 디코더 기반 LLM으로 생성한 임베딩에 듀얼 풀링을 적용하였을 때 유사한 효과가 나타나는지 검증하지 못하였다. 또한, LLM 임베딩을 활용하는 경우에도 mean 풀링에 결과를 고정하는 사례가 많아 다양한 풀링에서 S-BERT 임베딩과 LLM 임베딩의 상대적 성능을 체계적으로 비교한 연구가 부족하다.

본 연구는 BERTopic에서 S-BERT와 LLaMA 기반 문서 임베딩을 생성할 때, CLS, mean, max 그리고 듀얼 풀링이 토픽 일관성(Cv, Npmi)과 다양성에 미치는 효과를 비교하고자 한다. 이를 통해 토픽 모델링에서 임베딩 모델과 풀링 전략을 선택할 때 참고할 수 있는 실무적 재현 가능한 설계 지침을 제공하고자 한다.

연구 내용은 다음과 같다. 2장에서는 관련 선행연구와 임베딩에 대한 주요 개념을 기술하고, 3장에서는 문서 임베딩 생성을 포함한 연구 수행 방법에 대하여 상세하게 설명한다. 4장에서는 다양한 문서 임베딩 방법을 적용한 토픽 모델링 결과를 제시하고, 5장에서는 연구 요약 및 한계점, 향후 연구 방향에 관하여 기술한다.

Ⅱ. 이론적 배경

2.1 텍스트 임베딩(Text Embedding)

텍스트 임베딩이란 비정형 텍스트를 연속형 벡터로 표현하는 것을 의미하며, 의미적 유사성과 맥락을 수치적으로 다룰 수 있게 표현하는 방식이다. 초기의 임베딩은 TF-IDF[13]와 같이 단어의 출현빈도에 기반하여 문서를 벡터로 표현하

였고, Word2Vec[14]과 GloVe[15]가 대규모 말뭉치의 통계적 패턴을 학습해 단어 간 의미, 구문 관계를 포착하는 분산 표현을 확장하였다.

이후 BERT로 대표되는 사전학습 트랜스포머가 등장하면서 주변 문맥을 함께 모델링 하여 다의성을 반영할 수 있게 되었다. 특히, S-BERT[17]는 문맥 정보와 문서 임베딩은 쌍문장 대조 학습을 통하여 문장 수준 의미 공간을 직접 최적화함으로써 문서 유사도와 클러스터링 작업에 강점을 보였다. 한편 Transformer 언어 모델은 구조에 따라 입력 시퀀스를 한번에 인코딩하는 인코더 기반 모델과 이전 토큰을 조건으로 다음 토큰을 생성하는 디코더 기반 모델로 구분한다. BERT 계열은 양방향으로 인코딩하여 문맥표현을 하는 인코더 기반 모델이며, LLM 기반 모델은 자동 회귀 방식으로 토큰을 생성하는 디코더 모델 기반이다. BERT는 기본적으로 문서 임베딩을 직접 산출하지 않기 때문에, 여러 연구에서 출력 평균 풀링이나 CLS 토큰을 활용해 문서 수준의 고정 길이 벡터를 추출하는 방법이 제안되었다[16]. 풀링은 길이가 가변적인 시퀀스 표현을 고정 길이 벡터 형태로 집약하는 방법이다. 하지만 풀링은 생성 방식에 따라 문장 정보손실의 한계가 존재한다. 이에 최근 풀링 연구는 풀링을 결합해 그 한계점을 서로 보완하여 극복하고자 하였다. [11]은 CNN에 mean 풀링과 max 풀링의 평균을 내어 아랍어 감성분석 연구를 수행하여 모델의 안정성과 재현성을 제시하였다. [12]은 BERT의 기본 임베딩을 자동 에세이 채점 데이터에 적용하였고, 다양한 풀링의 방식을 조합하여 채점 성능을 향상하였다.

2.2 임베딩 기반 토픽 모델링의 발전

토픽 모델링은 대규모 비정형 문서를 주제 단위로 분석하는 방법이다. 전통적인 방법인 Latent Dirichlet Allocation(LDA)[18] 등의 토픽 모델링이 많이 사용되었지만 문맥을 고려하지 못하는 한계가 있다. 이러한 한계를 보완하기 위해 딥러닝을 활용한 BERTopic 기법이 등장하였으며, 사전학습 모델인 S-BERT로 문서 임베딩을 사용하여 대규모 데이터에 우수한 성능을 보였다[17].

[19]은 SCOPUS에서 ‘Language Model’ 키워드 관련 문헌을 수집하여 S-BERT와 BERTopic을 이용하여 14개의 토픽을 추출하고, 시계열, 국가 분포를 분석하였다. [20]는 바이든 공개 연설문 코퍼스를 이용하여 S-BERT, DistilBERT 등 임베딩을 BERTopic에 적용 및 비교하여 토픽 품질을 높였다. [21]는 S-BERT 임베딩과 BERTopic을 사용하여 전자상거래 영문과 중문 논문에 글로벌, 로컬 토픽 파악과 출현, 발전, 성숙, 쇠퇴로 이어지는 진화 패턴과 생애 주기를 탐구하였다.

최근에는 LLM 임베딩이 S-BERT보다 다양한 영역과 문맥 파악에 성능이 높다는 가능성이 제기되면서 LLM을 이용한 임베딩의 연구가 되고 있다. [8]은 다양한 LLM 임베딩을 모아 클러스터링 지표 중심으로 성능을 비교하였다. [9]은 LLM 임베딩이 텍스트 클러스터링에서 F1, ARI, HS 등 여러 지표 기준으로 전통임베딩 대비 우수함을 보여줌으로써 LLM 임베딩의 가능성을 보여주었다. [10]은 LLaMA, S-BERT, DistilBERT 등으로 문서 임베딩을 생성하여 BERTopic에 적용하였고, 대규모 코퍼스를 전처리 단계를 세분화하여 토픽 일관성, 토픽 다양성, 중요도의 성능을 비교하였다.

III. 연구방법

본 장은 연구의 절차와 이론을 설명하고자 한다. 연구 절차와 이론을 이해함으로써 연구의 이해를 돕고자 한다. 연구 절차는 그림 1과 같다.

3.1 문서 임베딩 방법

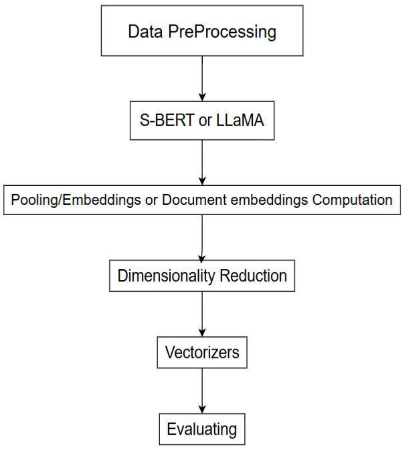

그림 1. 연구 절차

텍스트를 토크나이즈/패딩 한 뒤 Transformer 모델의 전방 추론으로 토큰별 은닉 상태 $h_{1:T} \in R^d$를 얻는다. 어태션 마스크로 패딩 토큰을 제외하고 풀링을 적용해 고정 길이$H \in R^{T \times d}$를 문서 임베딩 z($z \in R^d$)으로 만든다. d는 히든 차원이며, T는 시퀀스의 토큰수를 뜻한다 . 풀링은 크게 CLS, max, mean이 있으며, CLS 풀링 식은 아래와 같다.

$$z_{CLS} = h_1 \quad (1)$$

CLS 풀링은 최종 계산된 값인 첫 번째 토큰을 추출하여 사용한다. mean 풀링의 식은 아래와 같다.

$$z_{mean} = \frac{1}{N}\sum_{i=1}^{M} m_i h_i \quad (2)$$

N은 유효 토큰 수, M은 최대 시퀀스 길이,m_i는 i번째 토큰이 유효한지 나타내는 마스크을 뜻한다. mean 풀링은 트랜스포머 마지막 층이 내놓는 토큰 임베딩 행렬을, 패딩을 제외한 유효 토큰들만 골라 토큰 축으로 성분별 평균을 내서 길이 d 차원 벡터로 만드는 것을 뜻한다. max 풀링의 식은 아래와 같다.

$$z_{\max,j} = \max_{i:m_i=1} [h_i]_j \ , \ j = 1,\ldots,d \quad (3)$$

$z_{\max,j}$는 m_i=1인 유효 토큰들에 대해 각 차원 j에서의 최댓값을 취해 정의한다. 또한 j는 히든 차원 인덱스를 나타낸다. max 풀링은 전체 토큰 중 각 특징 차원에서 패딩을 제외한 유효 토큰들 중에 가장 큰 값을 선택하여 표현 벡터를 구성한다. 듀얼 풀링의 식은 다음과 같다.

$$z = [z_{\max} \parallel z_{mean}] \in R^{2d} \quad (4)$$

듀얼 풀링은 여러 층의 표현을 연결(concat) 하하여 하나의 더 긴 벡터를 만드는 방식이다. 만약 기본 차원이 d= 384일 때 최종 임베딩 차원은 2d = 768으로 늘려 토픽 모델에 많은 정보를 제공한다.

3.2 임베딩 모델

본 연구는 문서 임베딩 모델로 S-BERT의 all-MiniLM-L6-v2와 LLM의 Llama-3.2-1B를 사용한다. HuggingFace에서 제공되는 간편한 파이프라인 설정으로 사전학습 모델을 활용한다.

all-MiniLM-L6-v2는 자기 지도 대조 학습 목표를 사용하여 매우 큰 문장 단위 데이터셋에서 문장 임베딩 모델을 학습하는 것을 목표로 한다.

Llama3.2-1B는 다국어 대규모 언어 모델로 1B 크기로 사전학습되고 명령어 조정된 생성 모델 컬렉션이다. Llama3.2 모델은 에이전트 검색 및 요약 작업을 포함한 다국어 대화 사용 사례에 최적화되어있다.

표 1. 모델 환경

Embedding	Layers	Size	Vocabulary	Sequence length
all-MiniLM-L6-v2	6	384	30,522	256
Llama3.2-1B	16	2,048	128,256	131,072

3.3 BERTopic

본 연구는 트랜스포머 임베딩을 활용한 군집 기반 토픽 모델 BERTopic을 채택한다. BERTopic 하이퍼파라미터는 표 2로 고정하여 적용한다.

표 2. BERTopic 하이퍼파라미터

구성	파라미터	기본값
UMAP	n_neighbors	15
	n_components	5
	min_dist	0
	metric	'cosine'
	random_state	None
HDBSCAN	min_cluster_size	10
	metric	'euclidean'
	cluster_selection_method	'eom'
c-TF-IDF	top_n_words	10

본 연구의 파이프라인은 총 4단계로 요약된다. 첫째, 문장/문단 단위를 사전학습 트랜스포머로 벡터화하여 이후 단계의 입력으로 사용한다. 둘째, 임베딩을 차원 축소 기법 중 UMAP을 적용하여 지역/전역 구조를 보존하고 거리 기반 구분력을 확보한다. 셋째, 축소된 표현에 HDBSCAN을 적용해 가변 밀도와 노이즈를 처리하며 비관련 문서의 강제 배정을 방지한다[22]. 넷째, 각 군집을 하나의 문서로 간주해 c-TF-IDF로 대표 어휘를 산출하고, 토픽을 설명하는 핵심 키워드를 도출한다. 이러한 구성은 문맥 임베딩의 표현력과 밀도 기반 군집의 견고성을 결합함으로써 토픽 일관성과 해석 가능성을 높이는 것으로 보고된다[10].

3.4 평가지표

토픽 일관성에는 흔히 사용되는 Python Gensim 라이브러리를 활용한다. Gensim 라이브러리의 토픽 일관성 지표 중 Cv(Coherence Value)와 NPMI(Normalized Pointwise Mutual Information)를 채택한다. 또한 토픽의 다양성도 함께 확인한다.

토픽 내에서 상호 연관성이 높은 단어 쌍의 일관성을 정량화하기 위하여 NPMI을 채택한다. NPMI는 코퍼스나 설정이 달라져도 값의 해석과 비교가 쉽다는 장점이 있다. P(w_i), P(w_j)는 단어의 확률을 뜻하며, P((w_i,w_j)는 두 단어의 동시 확률을 뜻한다[23].

$$NPMI(w_i,w_j) = \frac{\log \frac{P(w_i,w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log P(w_i,w_j)+\epsilon} \quad (5)$$

Cv는 NPMI 점수로 구성한 벡터 간 코사인 유사도를 이용하여 한 주제로 잘 묶이는지 파악하고자 한다. 토픽의 상위 N 개 단어$(x_1,\cdots x_N)$를 자신 및 다른 표현 단어와의 NPMI 값으로 산출 후 생성된 모든 토픽 표현 단어 벡터들을 더하여 토픽 벡터를 구한다. Cv 값은 각 토픽 표현 단어 벡터와 해당 토픽 벡터 사이의 모든 코사인 유사도의 산술 평균으로 계산한다[23]. Cv의 식은 다음과 같으며, v_i 는 각 단어의 NPMI벡터, $\bar{v}$ 는 NPMI의 평균을 뜻한다.

$$Cv = \frac{1}{N}\sum_{i=1}^{N}\cos(v_i,\bar{v}) \quad (6)$$

토픽의 상위 단어 중 해당 단어가 그 토픽에서만 나타나는 비율을 측정하기 위해 토픽 다양성(Topic Diversity, TD)을 사용하였다. TD의 식은 다음과 같다.

$$TD = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{x_i \in t^{(k)}} I((x_i)) \quad (7)$$

K는 토픽의 수이며 T는 토픽의 상위 단어 수를 뜻한다. 지수함수 $I(x_i)$는 조건이 참이면 1, 거짓이면 0을 반환한다. TD는 0과 1 사이의 값을 가지며, 클수록 어휘 중복이 적어 더 다양한 것으로 해석한다[24].

Ⅳ. 연구 결과

4.1 데이터 수집 및 전처리

본 연구에서는 Hugging Face의 SetFit 공개 데이터셋(https://huggingface.co/SetFit)를 활용하였다. 각 데이터셋은 기본적으로 학습(train)과 평가(test) 데이터로 분리되어 제공되지만, 본 연구에서는 일반화 성능 평가를 위하여 이를 통합하였다. 통합 기준 원본의 규모는 BBC News 2,2

25개, 20Newsgroups(20NG) 18,846개, IMDB 50,000개이며, 전처리는 다음과 같이 수행하였다.

알파벳, 숫자, 밑줄, 공백, 특수 기호(물음표, 느낌표 등)를 제거한 후 소문자로 변환하여 토큰화하였다. 이후 spacy의 'en_core_web_sm'을 적용하여 단어를 표제어 변환하였다.

텍스트 임베딩의 품질을 높이기 위하여 토큰 수가 지나치게 적은 문서는 제외하였으며[25], 임곗값은 IMDB 기준으로 5개 미만으로 설정하였다. NLTK의 영어 불용어 목록을 적용한 후, BBC News 데이터는 'say', 'mr', 20Newsgroups는 'say', '–', 무의미한 반복 문자열('maxaxax....','mg9vg9vg9v...'), IMDB는 'br', 's'를 추가하여 단어를 제거하였다. 전처리 결과 최종 문서 수는 BBC News 데이터는 2,225개, 20Newsgroups는 18,100개 IMDB 데이터는 총 50,000개 문서로 구성하였다.

4.2 실험결과 분석

본 연구에서 제안한 구조는 S-BERT와 LLaMA으로 문서 임베딩을 추출한 뒤 CLS, mean, max 풀링을 단일/듀얼 풀링으로 조합하여 BERTopic에 적용하였다. 주요 하이퍼파라미터는 min_topic_size=10, nr_topics ∈{2, 4, …, 24}로 설정하였고, 군집 결과에서 이상치로 분류되는 토픽(-1)은 제외하여 유효 토픽 수만을 분석에 사용하였다. 이러한 토픽들을 Cv, NPMI, TD 지표를 활용해 평균과 최댓값을 비교하여 최적 풀링 구성을 제시한다. CLS 토큰이 없는 LLM은 CLS 풀링을 적용하였을 때 BOS 토큰으로 대체하였다. 임베딩의 조건은 표 3와 같이 정의하였다. T는 True를 의미하며, F는 False를 의미한다.

그림 2는 문서 임베딩 생성 방법에 따라 데이터별 평균 Cv를 비교한 결과이다. BBC 데이터에서 S-BERT는 0.6844, LLaMA는 0.6804로 CX(CLS-max) 풀링 방법이 가장 높은 Cv 값으로 나타났다. 20NG는 M(mean) 풀링 방법에서 S-BERT 0.5871, LLaMA 0.5493으로 가장 높은 값을 보였다. IMDB는 S-BERT는 X(max) 풀링, LLaMA는 M(mean) 풀링이 가장 높게 나왔으며, S-BERT는 0.5712, LLaMA는 0.6616으로 나타났다. 모델 별 최고 평균은 단일 풀링에서 관찰되지만, S-BERT와 LLaMA 둘 다 높은 풀링은 CM(CLS-mean) 풀링이다. mean 풀링을 보았을 때 대체로 LLaMA가 S-BERT보다 더 좋은 성능을 내고 있다.

표 3. 임베딩 조건표

임베딩 조건	정의
CLS =T, max = F , mean = F	C
CLS =F, max = T, mean = F	X
CLS =F, max = F , mean = T	M
CLS =T, max = F, mean = T	CM
CLS =T, max = T, mean = F	CX
CLS =F, max = T, mean = T	XM
CLS =T, max = T, mean = T	CXM

표 4와 같이 최댓값은 BBC, 20NG의 LLaMA에서 듀얼 풀링이 가장 높았고, IMDB, 20NG의 S-BERT는 mean 풀링 방법이 가장 높았다. 또한 듀얼 풀링으로 얻은 Cv 값들을 비교하면, 전반적으로 S-BERT가 LLaMA보다 더 높은 성능을 기록하였다.

표 4. 데이터별 S-BERT와 LLaMA의 최대 Cv와 풀링

		nr_topics	Pooling	max_Cv
BBC	S-BERT	16	CX	0.7893
	LLaMA	20	CX	0.7558
20NG	S-BERT	20	M	0.6502
	LLaMA	10	CM	0.6158
IMDB	S-BERT	4	M	0.6498
	LLaMA	6	M	0.6616

NPMI를 사용하여 상위 단어 쌍의 공동 출현 강도를 평가하였다. 그림 3은 데이터별 풀링 방법에 따라 평균 NPMI를 나타낸다. BBC 데이터에서 S-BERT는 CXM(CLS-max-mean) 방법으로 0.1593, LLaMA는 CX(CLS-max) 방법으로 0.1581이 나타나 풀링 방법에서 가장 좋은 성능을 보였다.

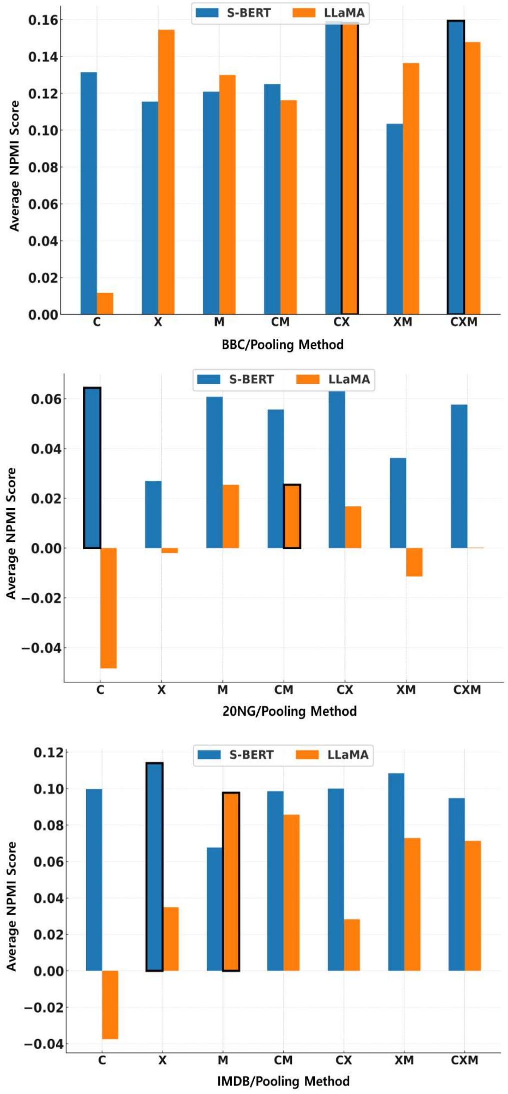

그림 2. S-BERT와 LLaMA의 평균 NPMI

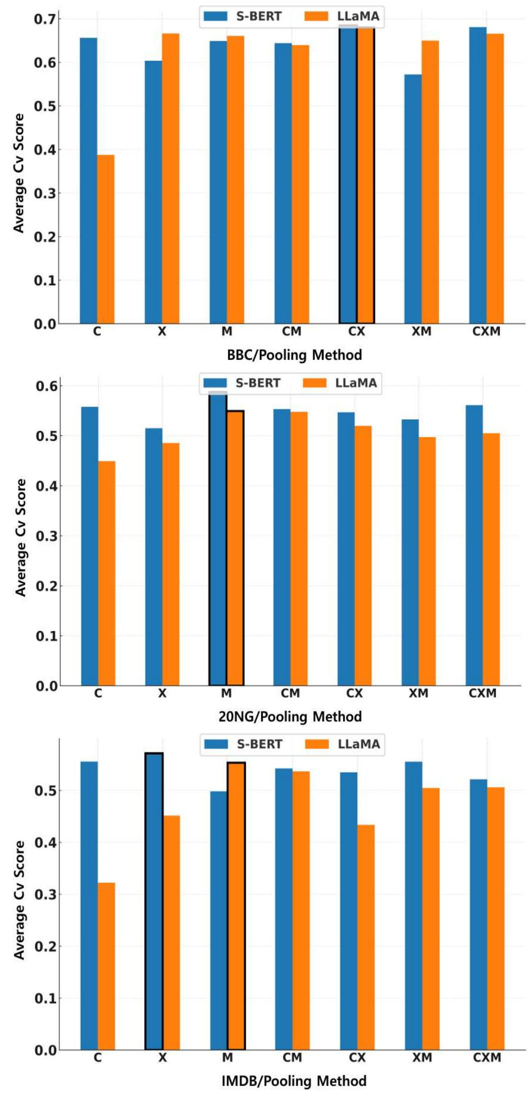

그림 3. S-BERT와 LLaMA의 평균 Cv

20NG에서는 S-BERT의 C(CLS) 풀링이 0.0643, LLaMA의 CM 풀링이 0.0254로 가장 높았다. LLaMA의 C(CLS), X(max) 풀링 평균 값이 음수였으나, 듀얼 풀링을 통해 양수로 개선되었다. IMDB는 S-BERT에서 0.1140로 X(max) 풀링이 높았고, LLaMA에서 0.0978로 M(mean) 풀링이 가장 높았다. mean 풀링에서 보았을 때 LLaMA 임베딩이 S-BERT보다 더 좋은 성능이 나오지만, 다양한 풀링 적용하였을 때 S-BERT가 LLaMA보다 더 좋은 성능을 이끌어냈다. 표 5에도 S-BERT가 LLaMA보다 비슷하거나 더 좋은 성능을 기록하였다.

표 5 데이터별 S-BERT와 LLaMA의 최대 NPMI와 풀링

	model	nr_topics	Pooling	max_NPMI
BBC	S-BERT	16	CX	0.2113
	LLaMA	4	CX	0.2221
20NG	S-BERT	22	CX	0.1246
	LLaMA	24	CXM	0.0829
IMDB	S-BERT	4	CXM	0.1638
	LLaMA	6	M	0.1549

Cv와 NPMI를 종합적으로 보면, S-BERT와 LLaMA는 풀링 방법에 대해 서로 다른 양상을 보였다. 인코더 모델인 S-BERT는 CLS나 max 풀링을 포함 및 사용하였을 때 mean 풀링보다 더 높은 성능을 내는 반면 디코더 모델인 LLaMA는 대체로 mean 풀링이 항상 상위권을 유지하며 비교적 안정적인 성능을 형성하였다.

Mean 풀링을 적용한 경우, LLaMA가 S-BERT보다 더 높은 Cv와 NPMI를 보이지만, 듀얼 풀링을 적용하면 오히려 S-BERT가 LLaMA보다 더 좋은 경향을 보였다. 표 4와 표 5에서도 최대 Cv, NPMI가 S-BERT 듀얼 풀링 조합에서 관찰되는 사례들을 확인할 수 있다. 또한 듀얼 풀링은 단일 풀링에 비해 약간 낮거나 비슷한 값을 보이는 경우도 있지만, 두 모델 모두에서 최적값에 근접한 수준의 성능을 유지한다는 점에서 의미가 있다.

이러한 결과는 mean 풀링이 임베딩에 가장 좋은 성능을 보장한다고 단정할 수 없음을 보여준다. 특히 언어 모델을 추가로 학습시키지 않더라도, 듀얼 풀링 조합만으로 S-BERT 임베딩이 Cv와 NPMI 측면에서 LLaMA로 생성한 임베딩보다 더 높은 성능을 달성할 수 있다. 이는 적절한 풀링 전략을 전제로 할 경우, S-BERT 임베딩이 LLM 임베딩을 대체할 수 있는 잠재력이 있음을 시사한다.

TD는 상위 단어들이 얼마나 다양하게 등장하는지를 평가하며 그림 4는 평균 TD를 나타낸다. BBC 데이터에서 S-BERT를 적용한 경우, 0.90에서 0.96 사이로 나타났으며, 20NG은 TD가 0.90에서 0.98로 분포되어 있다. IMDB는 다른 데이터보다 낮은 토픽 다양성을 보이며 0.86과 0.94 사이에 분포하고 있다.

LLaMA에서 C 풀링을 제외하였을 때, IMDB 제외한 두 데이터셋은 일정하게 0.85이상으로 나타났다,.IMDB는 0.65와 0.85 사이에 분포하고 M과 CM 풀링이 가장 높게 나왔다. 이러한 결과는 TD에서는 단일 풀링과 듀얼 풀링이 차이 크지 않음을 보여준다.

문서 임베딩을 생성시 단일 풀링과 듀얼 풀링의 자원 소모 및 연산 효율성을 비교함으로써, 듀얼 풀링 기법이 성능 향상에 비해 어느정도 연산 부담을 발생시키는지를 확인하고자 하였다. IMDB 기준 max 풀링과 mean-max 풀링을 기준으로 비교한 결과, 단일 풀링과 듀얼 풀링의 시간 차이는 크게 나지 않았다. 모든 실험은 동일한 설정에서 수행되었으며, 실험 환경은 S-BERT는 L4 GPU, LLaMA는 A100 GPU 기준으로 하였다. 표 6와 같이 S-BERT는 1분 41초, LLaMA는 57초의 차이로, 전체 파이프라인 관점에서 듀얼 풀링의 추가 연산 부담은 미미한 수준이었다. 하지만, S-BERT에 비하여 LLaMA 임베딩을 생성하는 데 더 높은 GPU를 사용함에도 불구하고, 약 4배의 시간 차이를 보였다.

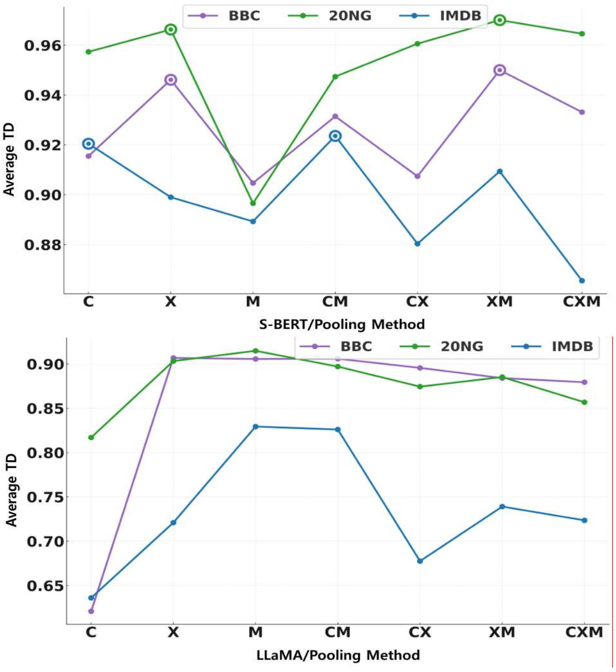

그림 4. S-BERT와 LLaMA의 평균TD 선그래프

표 6 단일 풀링과 듀얼 풀링 평균 소요시간(mm:ss)

	단일 풀링	듀얼 풀링
S-BERT	11.59	13.40
LLaMA	47.50	48.47

Ⅴ. 결 론

본 논문은 문서 임베딩 생성 방식에 따라 토픽 일관성과 다양성을 비교하고자 하였다. 이를 위해 S-BERT와 LLM 임베딩을 활용하여 문서 임베딩 생성을 위한 다양한 풀링 방법을 거친 후, BERTopic에 적용하여 토픽 모델링을 수행하였다. 모델 평가는 토픽 수는 2에서 24까지 토픽 일관성과 다양성으로 평가하였고, 산출한 지표의

평균과 최대 기준으로 비교하였다.

분석 결과, 단일 mean 풀링 관점에서는 LLaMA 임베딩이 S-BERT보다 상대적으로 높은 Cv와 NPMI를 보였으나, CLS, max 그리고 듀얼 풀링을 적용했을 때에는 세 데이터셋에서 S-BERT가 LLaMA 임베딩보다 토픽 일관성이 높게 나타났다. 특히, 듀얼 풀링을 S-BERT에 적용한 경우, 토픽 일관성이 LLaMA 임베딩에 mean 풀링을 적용하였을때보다 개선된 경우가 많아, S-BERT가 충분히 경쟁력 있는 대안이 될 수 있음을 확인하였다.

본 논문의 기여는 다음과 같다. 첫째, S-BERT와 LLaMA 임베딩을 mean 풀링에 한정하지 않고 다양한 풀링 전략의 관점에서 비교하였다. 동일 코퍼스와 동일 하이퍼파라미터 설정 하에서 문서 임베딩 풀링 방식만을 변경해 Cv, NPMI, TD를 측정함으로써, 풀링 전략에 따라 토픽 모델링 성능이 달라지는 풀링 민감도가 존재함을 실험적으로 보였다.

둘째, LLM 임베딩을 사용하지 않더라도, 풀링 전략을 조정함으로써 S-BERT 임베딩으로 충분히 경쟁력 있는 성능을 달성할 수 있음을 보였다. mean 풀링에서 LLaMA는 안정적이고 우수하였지만, 듀얼 풀링에서 S-BERT가 LLaMA보다 동등하거나 더 좋은 성능을 보였으며, 비용·지연시간 제약을 고려할 때 적절한 풀링 조합을 전제로 한 S-BERT 임베딩이 현실적인 대안이 될 수 있음을 보여준다. 셋째, 토픽 모델링 성능에서 문장을 시작하는 특수 토큰의 포함 여부가 중요함을 보였다. CLS 토큰이 없는 LLM에서도 CLS 계열 듀얼 풀링 구조가 일관되게 상위 성능을 보인다는 결과를 통해, 문장을 구분하는 시작 토큰의 존재가 성능 결정에 중요한 요소임을 실험적으로 확인하였다.

실무적 기여는 다음과 같다. 실무 환경에서는 LLM 임베딩에는 mean 풀링을 기본값으로 사용하고, LLM 사용이 어려운 경우에는 S-BERT 임베딩에 다양한 풀링을 적용하여 LLM 임베딩을 대체하거나 이에 필적하는 성능을 확보하는 전략을 권장한다. S-BERT와 LLM 모두에서 안정적인 토픽 일관성을 확보하고자 할 때에는 듀얼 풀링을 우선 적용해보고 데이터 특성에 따라 세부조정하는 전략을 추천한다.

그럼에도 본 연구에는 다음과 같은 한계가 있다. 첫째 다양한 LLM 모델을 사용하지 않아 다양한 결과를 포함하지 못한 한계가 있다. 둘째 토픽 모델의 하이퍼파라미터의 최적화를 체계적으로 진행하지 않았다. 하이퍼파라미터를 최적할 경우, 더 높은 수준의 일관성을 달성할 수 있다. 셋째, 본 연구는 자연어 처리 중 토픽 모델링 분야에 한정적으로 적용하였으며, 최근 주목받고 있는 LLM 기반 RAG(Retrieval-Augmented Generation)과 같은 프레임워크에서의 활용 가능성은 검증하지 못하였다. 향후 연구에서는 문서 임베딩 풀링 방식이 RAG와 같은 대규모 언어 모델 기반 지식 검색·생성 프레임워크 내에서 어떻게 작동하는지를 탐색하고자 한다. 특히 이러한 풀링 방식이 응답 품질과 검색 효율성에 미치는 영향을 실증적으로 분석할 계획이다.

REFERENCES

[1] J. Achiam et al., "GPT-4 Technical Report," arXiv:2303.08774, 2023.

[2] 장형운, 이민구, "디지털 커뮤니케이션 채널로서의 AI 교사보조로봇: 발명교육에서의 초등학생 인식 변화 탐색," *스마트미디어저널*, 제14권, 제9호, 43-54쪽, 2025년

[3] 전민규, 김남규, "텍스트 요약 품질 향상을 위한 의미적 사전학습 방법론," *Smart Media Journal*, 제12권, 제5호, 2023년

[4] 전민수 외, "초거대 언어 모델 기반 퀘적 예측을 활용한 사업 안전사고 예방," *Smart Media Journal*, 제14권, 제2호, 2025년

[5] 한무명초, "토픽 모델링과 네트워크 분석을 활용한 지역별 스마트관광 현황 분석: 뉴스기사를중심으로," *Smart Media Journal*, 제13권, 제9호, 2024년

[6] T. B. Brown et al., "Language Models Are Few-Shot Learners," arXiv:2005.14165, 2020.

[7] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning*

Research, 2020.

[8] I. Keraghel et al., “Beyond Words: A Comparative Analysis of LLM,” in *Advancing Intelligent Data Analysis (IDA)*, 2024.

[9] A. Petukhava, “Text Clustering with Large Language Model Embeddings,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 100–108, 2025.

[10] C. Yang and Y. Kim, “Enhancing Topic Coherence and Diversity in Document Embeddings Using LLMs: A Focus on BERTopic,” *Expert Systems with Applications*, vol. 281, art. 127517, 2025.

[11] M. Mhamed et al., “A Deep CNN Architecture with Novel Pooling Layer Applied to Two Sudanese Arabic Sentiment Data Sets,” *Journal of Information Science*, 2023.

[12] E. N. V. Boquio, “Beyond Canonical Fine-tuning: Leveraging Hybrid Multi-Layer Pooled Representations of BERT for Automated Essay Scorin,” in *ELRA Language Resource Association*, pp. 2285–2295, 2024.

[13] S. Robertson, “Understanding Inverse Document Frequency: On Theoretical Arguments for IDF,” *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[14] T. Mikolov et al., “Efficient Estimation of Word Representations in Vector Space,” *arXiv:1301.3781*, 2013.

[15] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” in *Proc. EMNLP*, 2014.

[16] Z. Zhang et al., “Is Neural Topic Modelling Better Than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics,” *arXiv:2204.09874*, 2022.

[17] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” in *Proc. EMNLP-IJCNLP*, pp. 3982–3992, 2019.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[19] W. Kang et al., “An Analysis of Research Trends on Language Model Using BERTopic,” in *Proc. 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, IEEE, 2023.

[20] R. M. Zain et al., “Revealing a Country’s Government Discourse Through BERT-Based Topic Modeling in the US Presidential Speeches,” in *Proc. 2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, IEEE, 2024.

[21] X. He, A. Fang, and D. Yu, “Multilingual Topic Evolution and Comparative Analysis for Electronic Commerce Research: A Combination of BERTopic and SBERT,” *Journal of Intelligent & Fuzzy Systems*, vol. 48, no. 1–2, pp. 1–22, 2025.

[22] M. Grootendorst, “BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure,” *arXiv:2203.05794*, 2022.

[23] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,” in *Proc. 8th ACM Int. Conf. on Web Search and Data Mining*, pp. 399–408, 2015.

[24] X. Wu, T. Nguyen, and A. T. Luu, “A Survey on Neural Topic Models: Methods, Applications and Challenges,” *Artificial Intelligence Review*, vol. 57, 2024.

[25] 홍인호, 김지은, 박지현, 주상훈, 지명인, “텍스트 임베딩과 키워드 네크워크를 통한 한국의 유네스코 세계유산 뉴스 동향 분석,” *Smart Media Journal*, 제13권, 제12호, 2024년

저 자 소 개

박지원(정회원)

2023년 계명대학교 통계학과 학사 졸업.

2025년 계명대학교 경영정보학과 석사 졸업.

<주관심분야 : 자연어처리, AI, 머신러닝>

양치복(정회원)

2016년 계명대학교 통계학과 학사 졸업.

2019년 계명대학교 경영정보학과 석사 졸업.

2025년 계명대학교 경영정보학과 박사 졸업.

<주관심분야 : 자연어처리, 머신러닝, 검색증강생성>

이충권(정회원)

1995년 계명대학교 경영정보학과 학사 졸업

1999년 Southeast Missouri State University MBA 졸업

2003년 University of Nebraska-Lincoln 경영학박사 졸업

2003-2006년 Georgia Southern University 조교수

<주관심분야 : 빅데이터, IT 인력>