Journal of Next-generation Convergence Information Services Technology Vol.15, No.1, February (2026), pp.83-96
ISSN 2384-101X(p) 2672-1163(e) *http://dx.doi.org/10.29056/jncist.2026.02.08*

추천 시스템에서 사전학습 언어모델 임베딩의 한계 : Semantic-Preference 불일치 현상 분석

Limitations of Pre-trained Language Model Embeddings in Recommender Systems: Analysis of Semantic-Preference Misalignment

한용희[1]*, 허진호[2]

Yong Hee Han[1]*, Jinho Hur[2]

요 약

사전학습 언어모델(PLM)의 성공에 힘입어 추천 시스템에서도 SBERT 등 PLM 임베딩 활용 연구가 활발히 진행되고 있으나, PLM이 추천 성능 향상에 항상 효과적인지에 대한 체계적 검증은 부족한 실정이다. 본 연구에서는 MovieLens-1M 데이터셋을 대상으로 SBERT(384차원) 임베딩의 효용성을 n=50 반복 실험을 통해 분석하였다. 실험 결과, 저밀도(5%) 환경에서 SBERT 임베딩이 오히려 추천 성능을 저해하는 것으로 나타났으며(ΔRMSE=+0.00046, p<0.0001), PCA 차원 축소(384→32차원) 적용 시 성능 저하가 4배 더 악화되어 과적합이 원인이 아님을 확인하였다. 이러한 현상은 PLM이 학습한 의미적 유사성 공간과 사용자 선호 공간 간의 불일치, 즉 의미-선호 불일치(Semantic-Preference Misalignment)로 설명된다. 본 연구는 제목과 장르만으로 구성된 제한적 텍스트 환경에서 미세조정 없이 사용되는 동결(Frozen) PLM의 효용성을 검증한 것으로, 풍부한 텍스트가 제공되거나 태스크 특화 미세조정이 적용되는 환경에서는 상이한 결과가 도출될 수 있다. 비용-효과 분석 결과, 추론 시간이 43% 증가하였음에도 정확도 개선이 관찰되지 않아, PLM 임베딩의 무비판적 도입에 대한 주의가 필요함을 시사한다.

핵심어 : 추천 시스템, 사전학습 언어모델, 의미-선호 불일치, 동결 임베딩, 협업 필터링

Abstract

Although the success of pre-trained language models (PLMs) in natural language processing has motivated active research on utilizing PLM embeddings such as SBERT in recommender systems, systematic verification of whether PLMs consistently enhance recommendation performance remains insufficient. This study analyzed the efficacy of SBERT (384-dimensional) embeddings on the MovieLens-1M dataset through n=50 repeated experiments. Experimental results revealed that SBERT

1 Department of Entrepreneurship and Small Business, Soongsil University, Seoul, Korea [Professor]
e-mail: amade@ssu.ac.kr (Corresponding author)
2 Department of Entrepreneurship and Small Business, Soongsil University, Seoul, Korea [Undergraduate Student]
e-mail: wlsgh9954@gmail.com

* 이 논문은 2025년도 교육부 및 서울특별시의 재원으로 서울RISE센터의 지원을 받아 수행된 서울시 지역혁신중심 대학지원체계(RISE)의 결과물입니다. (2025-RISE-01-020-04)

Received(December 18, 2025), Review Result(1st: January 13, 2026), Accepted(February 13, 2026), Published(February 28, 2026)

embeddings adversely affected recommendation performance in low-density (5%) environments (ΔRMSE=+0.00046, p<0.0001), and applying PCA dimensionality reduction (384→32 dimensions) exacerbated the performance degradation fourfold, thereby ruling out overfitting as the underlying cause. This phenomenon is attributed to the misalignment between the semantic similarity space learned by PLMs and the user preference space, termed Semantic-Preference Misalignment. This study examines the utility of frozen PLMs without fine-tuning in text-scarce environments consisting only of titles and genres; different results may be obtained in environments with richer textual information or task-specific fine-tuning. Cost-effectiveness analysis revealed a 43% increase in inference time without observable accuracy improvements, suggesting caution against the uncritical adoption of PLM embeddings in recommender systems.

Keyword : Recommender Systems, Pre-trained Language Models, Semantic-Preference Misalignment, Frozen Embeddings, Collaborative Filtering

1. 서론

자연어 처리(N에서 사전학습 언어모델(Pre-trained Language Model, PLM)은 BERT [1], SBERT [2]를 통해 성능을 끌어올리며 사전학습-미세조정 패러다임을 확립하였다. 이러한 흐름은 추천 시스템으로 확장되어, 아이템 텍스트를 PLM 임베딩으로 활용해 Cold-start 완화와 성능 향상을 기대하게 했다 [3-5]. 그러나 PLM 임베딩의 효과에 대한 체계적 검증은 제한적이며, 희소 데이터 편향과 반복 실험·유의성 검정 부족, 부정적 결과 보고의 부족이 지적된다.

본 연구는 미세조정 없이 동결된 PLM 임베딩을 특징으로만 사용하는 실무형 설정을 분석한다. 이는 추가 학습 없이 기존 추천 모델에 임베딩을 연결해 비용을 절감하는 방식이며, 현업에서 흔히 선택된다. 이러한 특징 기반 접근법은 널리 사용되지만, PLM 임베딩은 의미-선호 간극, BERT Anisotropy [6][7], 고차원 노이즈 문제로 성능 저해 가능성이 존재한다.

이에 본 연구에서는 세 가지 연구 질문을 제기한다. 첫째, PLM 임베딩(SBERT)이 추천 시스템 성능 향상에 항상 기여하는가? 둘째, PLM 임베딩이 부정적 효과를 나타내는 경우 그 원인이 과적합인가 아니면 의미-선호 간극(Semantic-Preference Gap)인가? 셋째, 차원 축소(PCA)가 PLM 임베딩의 효과를 개선하는가?

본 연구의 주요 기여는 다음과 같다. 첫째, 저밀도(5%) 환경에서 SBERT(384차원)가 통계적으로 유의하게 성능을 저해함을 n=50 반복 실험으로 입증하였다(p<0.0001, Cohen's d=0.90). 둘째, PCA 차원 축소가 오히려 성능을 악화시킴을 확인하여 단순한 과적합이 원인이 아님을 입증함으로써 과적합 가설을 기각하였다. 셋째, PLM의 의미 공간과 사용자 선호 공간의 불일치가 부정적 효과의 원인임을 분석하여 의미-선호 불일치(Semantic-Preference Misalignment) 현상을 규명하였다. 넷째, PLM 임베딩을 추천 시스템에 적용할 때의 주의사항 및 태스크 특화 적응의 필요성을 강조하는 실용적 지침을 제시하였다. 다섯째, 학습 및 추론 시간 측정을 통해 미세한 성능 차이를 위해 파이프라인 복잡도를 증가시키는 것의 비효율성을 정량적으로 제시하는 비용-효과 분석을 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 PLM을 활용한 추천 시스템 관련 선행 연구를 검토한

다. 3장에서는 데이터셋, 모델 아키텍처 및 실험 설계를 설명한다. 4장에서는 실험 결과를 제시하고, 5장에서는 결과의 해석과 원인 분석을 논의한다. 마지막으로 6장에서 결론을 맺는다.

2. 선행 연구

2.1 Semantic-Preference 정렬과 BERT Anisotropy

전통적인 추천 시스템은 Matrix Factorization [8] 등 협업 필터링을 기반으로 발전해 왔으며, PLM을 결합하는 연구는 (i)시퀀스 추천에 BERT 구조를 적용한 BERT4Rec [3]과 (ii) 텍스트 임베딩을 보조 정보로 결합하는 접근 [4][5][9]로 구분된다. 다만 BERT4Rec의 재현성 이슈 [10]와 관련 서베이 [11]를 고려할 때 효과의 일반성은 아직 불명확하다.

최근 연구들은 PLM의 의미 공간(Semantic Space)과 사용자 선호 공간(Preference Space) 간의 정렬(Alignment) 문제를 인식하고 이를 해결하려는 시도를 하고 있다. AlignRec [12]는 멀티모달 추천에서의 정렬 불일치(Misalignment) 문제를 체계적으로 분석하였으며, 기존 방법들이 멀티모달 정보를 단순히 보조 정보로 취급하여 ID 특징과의 의미적 간극(Semantic Gap)이 발생함을 지적하고 콘텐츠 내 정렬, 콘텐츠-ID 간 정렬, 사용자-아이템 간 정렬의 세 가지 정렬을 분리하여 학습하는 프레임워크를 제안하였다. 특히 주목할 만한 연구로 beeFormer [13]는 의미적 유사성과 상호작용 유사성 간의 간극을 직접적으로 다루었는데, 기존 Sentence Transformer 모델들이 상호작용 데이터의 숨겨진 패턴을 활용하지 않고 의미적 유사성만 예측하도록 학습된다는 점을 지적하며 상호작용 데이터로 Sentence Transformer를 직접 학습시키는 방법을 제안하였다.

한편, BERT 임베딩이 비등방적(Anisotropic) 특성을 보인다는 것은 Ethayarajh [14]에 의해 처음 지적되었으며, 모든 문장 임베딩이 좁은 원뿔(Cone) 형태로 뭉쳐 있어 코사인 유사도 기반의 비교가 의미 있는 차별성을 제공하지 못할 수 있다. 이 문제는 표현 퇴화(Representation Degeneration) 현상과 관련이 있으며, Gao et al. [15]은 자연어 생성 모델 학습에서 발생하는 표현 퇴화 문제를 분석하였다. 해결책으로 BERT-flow [6]는 임베딩 공간을 등방적(Isotropic) 가우시안 분포로 변환하는 학습을 제안하였고, BERT-whitening [7]은 후처리를 통한 정규화를 제안하였다. SimCSE [16]는 대조 학습(Contrastive Learning)을 통해 Anisotropy 문제를 완화하면서 문장 임베딩 성능을 크게 향상시켰다. 본 연구는 추천 시스템이라는 특정 다운스트림 태스크에서 BERT Anisotropy 및 Semantic-Preference Misalignment가 어떤 영향을 미치는지를 실증적으로 분석한다.

2.2 동결 PLM과 재현성 문제

PLM을 다운스트림 태스크에 적용하는 방식은 크게 Fine-tuning과 Feature Extraction(Frozen)으로 나뉜다. Howard & Ruder [17]의 ULMFiT 연구는 언어 모델 미세조정의 효과적인 방법론을 제시하

며 이 분야의 기초를 놓았다. Fine-tuning은 사전학습된 가중치의 일부 또는 전부를 업데이트하여 태스크에 적응시키는 방식으로 더 나은 성능을 달성할 수 있으나 계산 비용이 높고 데이터가 충분해야 한다. Feature Extraction은 사전학습된 가중치를 동결(Frozen)하고 새로운 층만 학습시키는 방식으로 적은 데이터와 계산 자원으로 빠른 프로토타이핑이 가능하나 태스크 적응이 제한적이다. 최근에는 Parameter-Efficient Fine-Tuning(PEFT)이 중간 지점으로 주목받고 있으며, 대표적으로 LoRA [18]는 저랭크 행렬을 통해 모델 파라미터의 극히 일부만 미세조정하면서 Full Fine-tuning에 근접한 성능을 달성할 수 있다. 본 연구는 현업에서 흔히 채택되는 Frozen Feature Extraction 방식의 한계를 실증적으로 분석하며, 추천 태스크에서 Plug-and-Play 방식의 PLM 적용이 효과적이지 않을 수 있음을 보여준다.

Dacrema et al. [19]은 최근 신경망 기반 추천 알고리즘들의 재현성 문제를 체계적으로 분석하였다. 18개의 알고리즘 중 7개만이 합리적인 노력으로 재현 가능하였으며, 재현 가능한 신경망 접근법 중 대부분이 개념적으로 단순한 방법에 의해 능가될 수 있었다. 이러한 발견은 'PLM을 추가하면 성능이 향상된다'는 일반적 가정에 대해서도 의문을 제기한다. 본 연구는 이러한 재현성 위기의 맥락에서, n=50 반복 실험과 엄밀한 통계적 검정을 통해 PLM 임베딩의 효용성을 검증하고자 한다.

3. 연구 방법

3.1 데이터셋 및 임베딩 생성

본 연구에서는 표준 공개 데이터셋인 MovieLens-1M을 사용하였다. MovieLens-1M 원본(6,040 사용자, 3,706 아이템, 밀도 약 4.5%)에서 고밀도 Subset을 구성하였으며, 반복적으로 평점 수가 임계값(τ_u, τ_i) 미만인 사용자와 아이템을 제거하여 수렴할 때까지 진행하였다. 본 연구에서 사용한 파라미터는 min_user_ratings=200, min_item_ratings=200, random_seed=42이며, 구성된 고밀도 Subset의 통계는 [표 1]과 같다.

[표 1] MovieLens-1M 고밀도 Subset 통계

[Table 1] Statistics of MovieLens-1M High-Density Subset

항목	값
원본 MovieLens-1M 밀도	~4.5%
Subset 사용자 수	2,945
Subset 아이템 수	2,514
Subset 밀도	11.18%
평점 범위	1 - 5

고밀도 Subset을 구성한 이유는 저밀도 환경에서 발생하는 극심한 분산을 줄이고 PLM 임베딩의 효과를 더 명확하게 관찰하기 위함이다. 그러나 이 필터링 과정에서 선택 편향(Selection Bias)이 발생할 수 있는데, 활발한 사용자와 인기 아이템만 포함되어 원본 MovieLens-1M의 대표성이 낮아지며 Cold-start 상황이 완전히 제외되었다. 따라서 본 연구의 결과는 “충분한 상호작용이 있는 Warm-start 환경”에 한정된다.

영화의 텍스트 정보(제목 + 장르 설명)를 Sentence-BERT 모델로 임베딩하였다. 모델은 paraphrase-multilingual-MiniLM-L12-v2(384차원)를 사용하였으며, 다국어 지원, 경량성, 표준성을 고려하여 선택하였다. 본 연구에서 영어 전용 모델 대신 다국어 모델을 선택한 것은 실무에서 범용적으로 적용 가능한 모델이 선호되며 본 연구는 이러한 “Plug-and-Play” 시나리오를 검증하기 위함이다. 이 가정의 타당성을 검증하기 위해 영어 전용 모델(all-MiniLM-L6-v2)과의 비교 실험도 수행하였다. 비교를 위해 Genre Multi-hot 인코딩(18차원)도 함께 실험하였다.

3.2 모델 아키텍처

PLM 임베딩의 효과를 검증하기 위한 대조군으로서 PureNCF를 구성하였다. 이 모델은 텍스트의 의미적 정보를 배제하고 상호작용 정보만을 활용하는 NCF(Neural Collaborative Filtering) [20] 아키텍처로, 순수한 협업 필터링 공간에서 동작한다. 손실 함수는 $L = \sum(R_{UV} - \widehat{R_{UV}})^2$이며, $\widehat{R_{UV}} = f(E_U \oplus E_V)$로 정의된다. 여기서 E_U는 사용자 잠재 벡터, E_V는 아이템 잠재 벡터, f는 비선형 변환을 수행하는 MLP이다.

HybridNCF는 PureNCF의 Preference Space에 PLM이 생성한 Semantic Space를 결합한 모델이다. $\widehat{R_{UV}} = f(E_U \oplus E_V \oplus S_V)$로 정의되며, S_V는 아이템 V의 SBERT 임베딩(384차원)이다. 본 연구의 핵심 질문은 “이 두 공간이 정렬(Align)되어 있는가?”이며, 두 공간이 불일치(Misalignment)하면 S_V가 노이즈로 작용할 위험이 존재한다.

3.3 실험 설계 및 평가

원본 고밀도 데이터(11.18%)에서 랜덤 샘플링을 통해 다양한 밀도 수준을 시뮬레이션하였다. 밀도 레벨은 5%, 10%, 11%의 세 수준으로 설정하였으며, 반복 횟수는 SBERT 실험의 경우 n=50, Genre 실험의 경우 n=2(탐색적)로 설정하였다.

SBERT 임베딩의 부정적 효과 원인을 규명하기 위해 추가 Ablation 실험(n=20)을 수행하였다. SBERT_384(원본), SBERT_PCA_32(분산 유지율 51%), SBERT_PCA_64(분산 유지율 73%), Random_384, Random_32 조건을 비교하였으며, 검증 가설은 “SBERT의 부정적 효과가 과적합 때문이라면, 차원 축소 시 성능이 개선될 것이다”이다.

PLM 임베딩이 추천 성능에 미치는 영향의 통계적 유의성을 검증하기 위해 다층적 분석을 수행하였다. 1단계 모수 검정에서는 n=50회 반복 실험에서 수집한 ΔRMSE 표본에 대해 Paired t-test를 적용하였다. 2단계에서는 Cohen's d로 효과 크기를 정량화하고 Bootstrap 신뢰구간(B=10,000)으로 불확실성을 평가하였다. 3단계에서는 Shapiro-Wilk 검정으로 정규성 가정을 확인하고 Wilcoxon signed-rank 검정으로 강건성을 분석하였다.

평가 지표로는 RMSE(Root Mean Squared Error)를 주 지표로 사용하였다. ΔRMSE는 HybridNCF RMSE에서 PureNCF RMSE를 뺀 값으로 정의하며, ΔRMSE<0이면 Side Information이 도움이 되고 ΔRMSE>0이면 해로운 것으로 해석한다.

4. 실험 결과

4.1 주 실험 결과

본 분석에는 다음과 같은 방법론적 한계가 존재한다. SBERT 실험(n=50)은 통계적 검정력이 높아 신뢰할 수 있으나, Genre 실험(n=2)은 매우 낮은 검정력으로 인해 결론의 신뢰도가 제한적이다. 따라서 본 절은 향후 연구 방향 제시를 위한 탐색적 관찰로서 확정적 결론 도출에는 한계가 있다. [표 2]는 Side Information 유형별 성능 비교 결과를 보여준다.

[표 2] Side Information 유형별 성능 비교 (탐색적 분석)

[Table 2] Performance Comparison by Side Information Type (Exploratory Analysis)

Side Info	차원	밀도	ΔRMSE	효과	반복	통계적 유의성
Genre	18	5%	-0.0016	개선 경향	n=2	미검증
Genre	18	10%	-0.0024	개선 경향	n=2	미검증
Genre	18	11%	-0.0010	개선 경향	n=2	미검증
SBERT	384	5%	+0.00046	부정적	n=50	p<0.0001
SBERT	384	10%	+0.00006	중립	n=50	p=0.578
SBERT	384	11%	-0.00006	중립	n=50	p=0.609

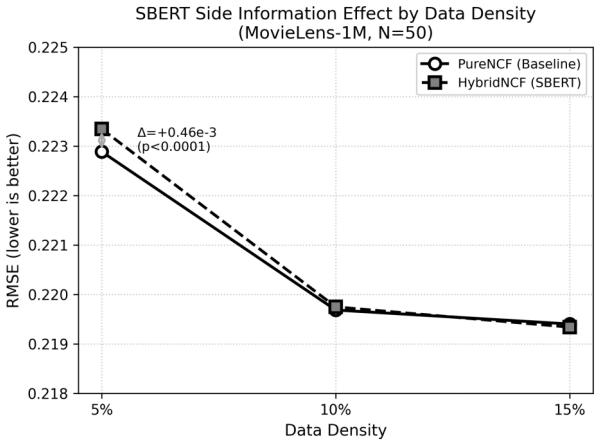
또한, [표 3]은 SBERT 실험(n=50)의 상세 통계 분석 결과를 보여준다.

[표 3] SBERT 실험 통계 분석 결과 (n=50)

[Table 3] Statistical Analysis Results of SBERT Experiments (n=50)

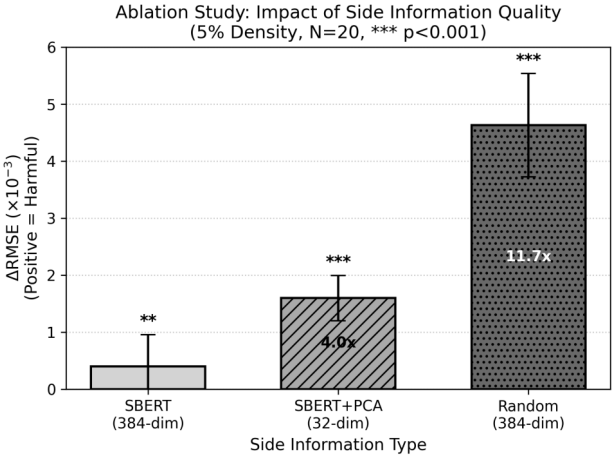
밀도	Pure	Hybrid	ΔRMSE	Bootstrap 95% CI	t-stat	p-value	Cohen's d	해석
5%	0.2229	0.2233	+0.00046	[+0.00032, +0.00060]	+6.33	7.17e-08	+0.90	부정적(유의)
10%	0.2197	0.2197	+0.00006	[-0.00015, +0.00028]	+0.56	0.578	+0.08	중립
11%	0.2194	0.2193	-0.00006	[-0.00030, +0.00018]	-0.52	0.609	-0.07	중립

[표 3]에서 확인할 수 있듯이, 5% 밀도에서는 p=7.17e-08 < 0.0001이며 Cohen's d = 0.90으로 큰 효과 크기를 보였다. 95% CI가 0을 포함하지 않아 'SBERT가 저밀도에서 부정적 효과를 나타낸다"는 결론이 통계적으로 지지된다. 본 결과는 min_ratings=200 필터링을 거친 고밀도 Subset에서 도출되었으므로, 본 연구에서의 "저밀도"는 Cold-start가 아닌 시뮬레이션 저밀도에 해당한다. [그림 1]은 이러한 밀도별 RMSE 비교를 시각화한 것으로, 5% 밀도에서 HybridNCF(SBERT)가 PureNCF(Baseline)보다 유의하게 높은 RMSE를 보이며(ΔRMSE=+0.46×10⁻³, p<0.0001), 10-11% 밀도에서는 유의한 차이가 관찰되지 않았다.



[그림 1] 밀도별 SBERT Side Information 효과

[Fig. 1] Effect of SBERT Side Information by Density



[그림 2] Bootstrap 95% 신뢰구간

[Fig. 2] Bootstrap 95% Confidence Intervals by Density

통계적 가정 검증 결과, Shapiro-Wilk 정규성 검정에서 모든 밀도 조건에서 정규성이 유지되었다 (5%: p=0.3884, 10%: p=0.4275, 11%: p=0.0584). 모수 검정(t-test)과 비모수 검정(Wilcoxon)의 결론이 일관되어 검정 방법 선택에 따른 결론 변동이 제한적임을 확인하였다. [그림 2]는 n=50 반복 실험에서 B=10,000 Bootstrap 리샘플링을 통해 산출한 95% 신뢰구간을 시각화한 것으로, 5% 밀도 조건에서만 신뢰구간이 0을 포함하지 않아 통계적으로 유의한 부정적 효과가 확인된다.

4.2 Ablation 분석

SBERT 임베딩의 부정적 효과 원인을 규명하기 위한 Ablation 실험 결과는 [표 4]와 같다.

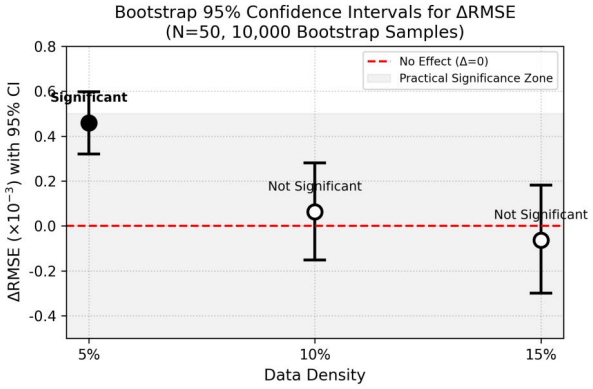
[표 4] Ablation 실험 결과 (n=20)

[Table 4] Ablation Experiment Results (n=20)

조건	5% ΔRMSE	10% ΔRMSE	11% ΔRMSE	해석
SBERT_384	+0.00040*	+0.00041	+0.00002	5%에서만 유의하게 부정적
SBERT_PCA_32	+0.00160*	+0.00088*	+0.00071*	4배 더 부정적
SBERT_PCA_64	+0.00186*	+0.00102*	+0.00083*	4.6배 더 부정적
Random_384	+0.00463*	+0.00531*	+0.00520*	가장 부정적(순수 노이즈)
Random_32	+0.00308*	+0.00337*	+0.00317*	Random도 부정적

*p<0.05 (Paired t-test)

PCA는 훈련 아이템에서 적합(fit)한 후 전체 아이템에 변환(transform)을 적용하였다. 본 실험에서는 단순 PCA(whiten=False)를 사용하였으며, Whitening을 적용한 PCA나 BERT-whitening과 같은 후처리 기법은 검증하지 않았다.



[그림 3] Ablation 연구: Side Information 품질별 성능 저하 비교 (5% 밀도, n=20).

[Fig. 3] Ablation Study: Performance Degradation Comparison by Side Information Quality (5% Density, n=20).

[표 4]에서 확인할 수 있는 주요 발견은 다음과 같다. 첫째, PCA 차원 축소(384→32)가 오히려 성능을 4배 악화시켜 과적합 가설이 기각되었다. 둘째, PCA가 보존하는 고분산 축이 추천 태스크와 무관한 노이즈일 가능성을 보여주는 노이즈 농축 현상이 관찰되었다. 셋째, SBERT가 Random Noise보다 우수한 성능을 보여(+0.00040 vs. +0.00463) 일부 유용한 정보가 포함되어 있음을 확인하였다. 이러한 결과는 [그림 3]에 시각화하였다.

4.3 비용-효과 분석

비용-효과(ROI) 관점에서 PureNCF와 HybridNCF(SBERT)의 학습/추론 시간을 측정하였으며(n=10), 그 결과는 [표 5]와 같다.

[표 5] 학습 및 추론 시간 비교 (n=10 평균)

[Table 5] Training and Inference Time Comparison (n=10 Average)

밀도	모델	RMSE	학습 시간(s)	추론 시간(s)	ΔRMSE
5%	PureNCF	0.2235	27.97	0.47	-
5%	HybridNCF	0.2244	27.22	0.67	+0.0008
10%	PureNCF	0.2194	54.53	1.17	-
10%	HybridNCF	0.2197	54.58	1.17	+0.0003
11%	PureNCF	0.2189	61.72	1.54	-
11%	HybridNCF	0.2189	61.54	1.08	≈0

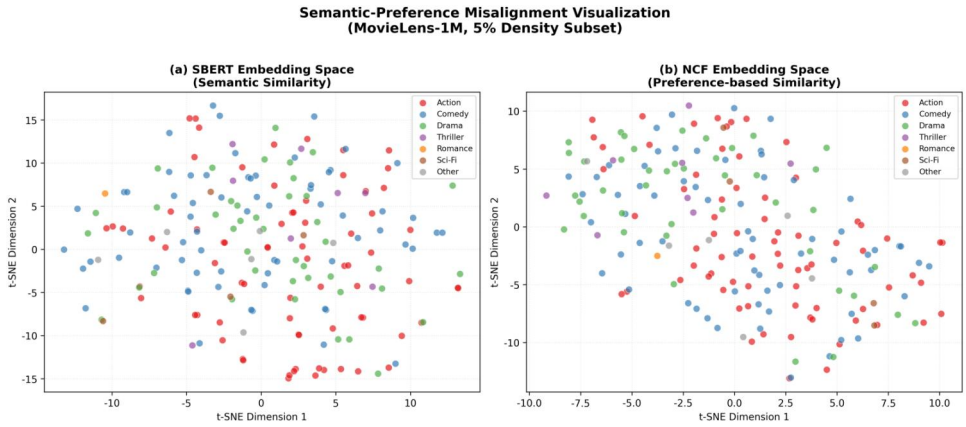
[표 5]에서 확인할 수 있듯이, 학습 시간 측면에서 PureNCF와 HybridNCF가 거의 동일하며 이는 SBERT 임베딩이 사전 계산되어 오버헤드가 미미하기 때문이다. 추론 시간 측면에서 5% 밀도에서 HybridNCF가 약 1.4배 느렸다(0.67s vs. 0.47s). ROI 분석 결과, RMSE가 0.08% 악화되면서 추론 시간은 43% 증가하여 SBERT Side Information의 비용-효과가 부정적인 것으로 나타났다.

5. 논의

5.1 결과 해석

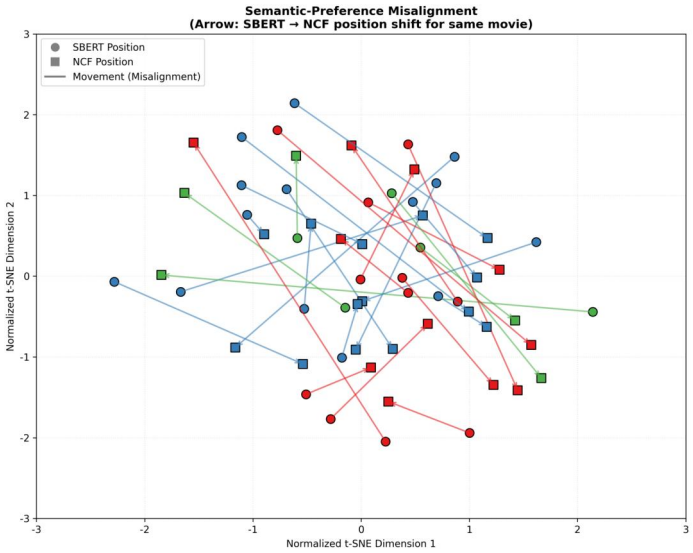
본 연구의 핵심 발견은 PLM의 의미 공간(Semantic Space)과 사용자 선호 공간(Preference Space)이 정렬되지 않는다는 것이다. Semantic Space(SBERT)는 줄거리, 장르, 키워드 등 텍스트 기반 유사성을 반영하는 반면, Preference Space(추천)는 개인 취향, 맥락, 사회적 요인 등 행동 기반 유사성을 반영한다. 핵심적인 Gap은 의미적으로 유사한 영화라도 사용자 선호는 다를 수 있다는 점이다. 예를 들어, 같은 SF 장르라도 하드 SF와 소프트 SF에 대한 선호는 극명하게 갈릴 수 있다.

t-SNE 시각화를 통해 이 불일치를 시각적으로 확인하였으며, 그 결과는 [그림 4]와 같다. SBERT 임베딩 공간에서 동일 장르 영화가 비교적 가깝게 위치하는 반면, NCF 학습 공간에서는 사용자 선호 패턴에 따라 장르와 무관하게 재배치된다. 예를 들어 'The Matrix"는 SBERT 공간에서 'Independence Day'와 가까우나, NCF 공간에서는 'Blade Runner' 등 사이버펑크 영화와 가까워진다. 또한, [그림 5]는 동일 영화들이 두 임베딩 공간 사이에서 어떻게 이동하는지를 시각화하여, 의미-선호 불일치의 양상을 더욱 명확하게 보여준다.



[그림 4] SBERT 공간과 NCF 학습 공간의 t-SNE 시각화 비교.

[Fig. 4] t-SNE Visualization Comparison of SBERT Space and NCF Learned Space.



[그림 5] 동일 영화의 두 임베딩 공간 간 위치 이동 시각화.

[Fig. 5] Visualization of Position Shifts Between Two Embedding Spaces for the Same Movies.

정량적 정렬도 분석 결과, kNN Overlap@10=0.119로 임계값 0.3을 크게 밑돌았으며, Similarity Correlation = -0.020으로 거의 0에 가까웠다. 이는 SBERT 임베딩 공간과 NCF 학습 공간의 이웃 구조가 근본적으로 다름을 확인한다. 또한 다양한 정규화 기법(L2, StandardScaler, Temperature)을 적용한 ablation 실험에서도 성능이 개선되지 않아 문제의 본질이 스케일 불일치가 아님을 확인하였다. 영어 전용 SBERT 모델(all-MiniLM-L6-v2)과의 비교 실험(n=10)에서도 5% 밀도에서 PLM이 성능 향상에 기여하지 못하였다(ΔRMSE$\approx$0, p=0.95). 이는 모델 선택이 아닌 Semantic-Preference Misalignment가 근본 원인임을 확인한다.

Ablation 실험(Table 4)에서 PCA가 성능을 악화시킨 이유는 BERT 임베딩의 비등방성으로 유용한 저분산 정보가 제거되고 노이즈가 농축되기 때문이다. 또한 저밀도에서 Side Information이 부정적 효과를 나타낸다는 발견은 기존 예상과 상반된다. 본 연구의 "저밀도(5%)"는 min_ratings$\geq$200인 활발한 사용자를 대상으로 한 시뮬레이션 저밀도로서 실제 Cold-start 환경과는 구별된다. 따라서 PLM이 Cold-start를 해결하지 못한다"는 주장은 본 연구에서 검증되지 않았다.

본 연구의 결과는 Side Information의 품질과 특성에 따라 효과가 달라짐을 시사한다. 전문가 레이블(Genre, Tag)은 저차원이고 정제되어 있어 추천 태스크와의 정렬도가 높은 반면, Frozen PLM 임베딩은 고차원이고 범용적이어서 정렬도가 낮다. 따라서 저밀도에서 Side Information이 도움이 된다는 명제는 Side Information의 품질이 높을 때만 성립한다.

5.2 실무적 함의 및 한계

통계적 유의성과 실용적 유의성의 구분이 필요하다. 5% 밀도에서 SBERT의 ΔRMSE가 +0.00046(0.21% 악화)이며, p<0.0001로 통계적으로 매우 유의하고 Cohen's d=0.90으로 큰 효과 크기를 나타낸다. 그러나 Cohen's d=0.90은 n=50의 대규모 반복 실험에서 분산이 현저히 감소하여 절대적으로 미미한 차이도 표준화된 지표에서 큰 값을 나타낼 수 있음에 주의해야 한다. 본 연구의 핵심적인 발견은 효과의 크기가 아닌 효과의 방향(Direction)에 있다. 이러한 분석에 기반한 PLM 임베딩 적용 지침은 [표 6]과 같다.

[표 6] PLM 임베딩 적용 지침

[Table 6] Guidelines for PLM Embedding Application

상황	권장 사항	근거
저밀도(<10%)	PLM 임베딩 적용 전 신중한 검증	부정적 효과 가능성
중고밀도($\geq$10%)	PLM 임베딩 효과 미미	협업 신호로 충분
PLM 사용 필요	태스크 특화 적응(Fine-tuning) 권장	의미-선호 정렬 필요
차원 축소	단순 PCA 적용 지양	노이즈 농축 위험

PLM 임베딩 도입 전 검증 방안으로는 다음의 세 가지를 고려할 수 있다. 첫째, PLM 임베딩만으로 평점을 예측하는 간단한 회귀 분석을 통해 Side Information의 단독 예측력을 평가하는 방법이 있다. 둘째, ID 임베딩과 PLM 임베딩 간의 kNN overlap@10을 계산하여 공간 정렬도를 확인할 수 있다. 셋째, Gated Fusion을 사용할 경우 PLM 가중치가 0으로 수렴하는지 관찰하는 것이 유용하다.

Semantic-Preference Misalignment 문제를 해결하기 위한 잠재적 접근법으로는 Full Fine-tuning, LoRA, Adapter Layer, Contrastive Learning 등이 있다. 특히 beeFormer[15]처럼 상호작용 데이터로 Sentence Transformer를 직접 학습시키는 정렬 학습이 유망하다. Gated Fusion ($\hat{y} = g \cdot e_{PLM} + (1-g) \cdot e_{ID}$)을 통해 PLM 기여도를 자동 조절할 수도 있다.

본 연구의 주요 한계점은 다음과 같다. 첫째, 텍스트 입력이 제목과 장르만으로 빈약하여 Amazon Review나 IMDB 등 풍부한 텍스트 환경에서의 재검증이 필요하다. 둘째, Genre 실험이 n=2로 불균형하여 $n \geq 30$으로 재실험이 권장된다. 셋째, Warm-start 환경에 한정되어 Cold-start 시뮬레이션 실험이 필요하다. 넷째, SBERT 외의 다른 PLM 모델에 대한 추가 검증이 필요하며 Fine-tuning, LoRA, Adapter 등의 해결책은 미검증 상태이다. 다섯째, MovieLens-1M 단일 데이터셋만 사용하여 다른 도메인에서의 검증이 필요하다.

6. 결론

본 연구에서는 추천 시스템에서 사전학습 언어모델(PLM) 임베딩의 효용성을 실증적으로 분석하였다.

주요 결론은 다음과 같다. 첫째, 저밀도(5%) 환경에서 SBERT(384D)가 통계적으로 유의하게 성능에 부정적 영향을 미치는 것으로 관찰되었다(p<0.0001, Cohen's d=0.90). 단, 효과의 절대적 크기(ΔRMSE=+0.00046, 0.21% 악화)는 실용적으로 무시할 수 있는 수준이다. 핵심 발견은 효과의 방향이 기존의 “Side Information이 저밀도에서 도움이 된다”는 가정과 상반된다는 점이다.

둘째, PCA 차원 축소가 오히려 성능을 4배 악화시켜 과적합 가설이 기각되었다([표 4] 참조). 단순한 과적합(Curse of Dimensionality)이 원인이 아님을 확인하였다.

셋째, PLM의 의미 공간과 사용자 선호 공간이 정렬되지 않아 PLM 임베딩이 추천 태스크에서 효과적으로 활용되지 못하는 Semantic-Preference Misalignment 현상을 규명하였다([그림 4], [그림 5] 참조).

넷째, 실용적 함의로서 PLM 임베딩을 추천 시스템에 Plug-and-Play 방식으로 적용할 경우 기대했던 긍정적 효과가 관찰되지 않을 수 있으며, 태스크 특화 적응 없이는 효과가 미미하거나 방향이 반대로 나타날 수 있음을 확인하였다([표 6] 참조).

다섯째, 비용-효과(ROI) 관점에서 SBERT Side Information 추가는 RMSE를 0.08% 악화시키면서

추론 시간은 43% 증가시켰다(5% 밀도 기준, [표 5] 참조). 현업에서 미세한 성능 개선을 위해 파이프라인 복잡도를 증가시키는 것은 비효율적일 수 있다.

본 연구의 결론은 빈약한 텍스트(제목+장르) + Frozen SBERT + 단순 Concatenation이라는 특정 조건에서 도출되었음에 유의해야 한다. 풍부한 텍스트가 제공되거나 Fine-tuning/Adapter를 통해 PLM을 태스크에 적응시킨 경우에는 상이한 결과가 관찰될 가능성이 있다. 그럼에도 불구하고 본 연구는 PLM 임베딩의 무비판적 도입에 대한 경고를 제시하며, Plug-and-Play 방식 적용 전 신중한 검증의 필요성을 시사한다.

References

[1] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), June 2-7, 2019, Minneapolis, MN, USA, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[2] N. Reimers, I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-networks”, 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), November 3-7, 2019, Hong Kong, China, pp. 3982-3992, doi: 10.18653/v1/D19-1410.

[3] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, “BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer”, 28th ACM International Conference on Information and Knowledge Management (CIKM), November 3-7, 2019, Beijing, China, pp. 1441-1450, doi: 10.1145/3357384.3357895.

[4] Q. Zhang, J. Li, Q. Jia, C. Wang, J. Zhu, Z. Wang, X. He, “UNBERT: User-news matching BERT for news recommendation”, 30th International Joint Conference on Artificial Intelligence (IJCAI-21), August 19-26, 2021, Montreal, Canada (Virtual Event), pp. 3356-3362, doi: 10.24963/ijcai.2021/462.

[5] C. Wu, F. Wu, T. Qi, Y. Huang, “Empowering news recommendation with pre-trained language models”, 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), July 11-15, 2021, Online (Virtual Event, originally Montreal, Canada), pp. 1652-1656, doi: 10.1145/3404835.3463069.

[6] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, L. Li, “On the sentence embeddings from pre-trained language models”, 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), November 16-20, 2020, Online, pp. 9119-9130, doi: 10.18653/v1/2020.emnlp-main.733.

[7] J. Su, J. Cao, W. Liu, Y. Ou, “Whitening sentence representations for better semantics and faster retrieval”, arXiv preprint arXiv:2103.15316, March 2021, doi: 10.48550/arXiv.2103.15316.

[8] Y. Koren, R. Bell, C. Volinsky, “Matrix factorization techniques for recommender systems”, Computer, vol. 42, no. 8, August 2009, pp. 30-37, doi: 10.1109/MC.2009.263.

[9] M. Ibanez, D. Vial, J. A. Pino, “V-BERT4Rec: Enhanced sequential recommendation with multi-modal visual information”, Multimedia Tools and Applications, vol. 83, no. 1, January 2024, pp. 1-22, doi:

10.1007/s11042-024-19277-7.

[10] A. V. Petrov, C. Macdonald, "A systematic review and replicability study of BERT4Rec for sequential recommendation", 16th ACM Conference on Recommender Systems (RecSys), September 18-23, 2022, Seattle, WA, USA, pp. 436-447, doi: 10.1145/3523227.3548487.

[11] P. Liu, L. Wang, J. Cao, W. Shi, J. Ye, Y. Zhang, "Pre-train, prompt, and recommendation: A comprehensive survey of language modelling paradigm adaptations in recommender systems", arXiv preprint arXiv:2302.03735, February 2023, doi: 10.1162/tacl_a_00619.

[12] Z. Liu, Y. Ma, Y. Ouyang, W. Xiong, "An aligning and training framework for multimodal recommendations", arXiv preprint arXiv:2403.12384, March 2024, doi: 10.48550/arXiv.2403.12384.

[13] J. Kasanicky, P. Peska, L. Holena, "beeFormer: Bridging the gap between semantic and interaction similarity in recommender systems", arXiv preprint arXiv:2409.10309, September 2024, doi: 10.1145/3640457.3691707.

[14] K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings", 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), November 3-7, 2019, Hong Kong, China, pp. 55-65, doi: 10.18653/v1/D19-1006.

[15] J. Gao, T. He, X. Zhou, Z. Shao, B. Qin, "Representation degeneration problem in training natural language generation models", arXiv preprint arXiv:1907.12009, July 2019, doi: 10.48550/arXiv.1907.12009.

[16] T. Gao, X. Yao, D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings", 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), November 7-11, 2021, Online and Punta Cana, Dominican Republic, pp. 6894-6910, doi: 10.18653/v1/2021.emnlp-main.552.

[17] J. Howard, S. Ruder, "Universal language model fine-tuning for text classification", 56th Annual Meeting of the Association for Computational Linguistics (ACL), July 15-20, 2018, Melbourne, Australia, pp. 328-339, doi: 10.18653/v1/P18-1031.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, "LoRA: Low-rank adaptation of large language models", arXiv preprint arXiv:2106.09685, June 2021, doi: 10.48550/arXiv.2106.09685.

[19] M. F. Dacrema, P. Cremonesi, D. Jannach, "Are we really making much progress? A worrying analysis of recent neural recommendation approaches", 13th ACM Conference on Recommender Systems (RecSys), September 16-20, 2019, Copenhagen, Denmark, pp. 101-109, doi: 10.1145/3298689.3347058.

[20] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T. S. Chua, "Neural collaborative filtering", 26th International Conference on World Wide Web (WWW), April 3-7, 2017, Perth, Australia, pp. 173-182, doi: 10.1145/3038912.3052569.