JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 10, pp. 2835-2850, Oct. 2026

Check for updates

온라인 리뷰 유용성 예측을 위한 도메인 일반화 알고리즘 비교 연구

김 유 진[1] · 박 정 은[2*]
[1]한양대학교 휴먼컴퓨터인터랙션학과 석사과정
[2]한양대학교 휴먼컴퓨터인터랙션학과 조교수

Comparative Analysis of Domain Generalization Algorithms for Prediction of Online Review Helpfulness

Yujin Gim[1] · Jeongeun Park[2*]

[1]Master's Course, Department of Human-Computer Interaction, Hanyang University, Ansan 15588, Korea
[2]Assistant Professor, Department of Human-Computer Interaction, Hanyang University, Ansan 15588, Korea

[요 약]

기존 리뷰 유용성 예측 연구는 학습 데이터와 테스트 데이터가 동일한 분포를 따른다는 독립항등분포 가정에 기반하고 있다. 그러나, 실제 온라인 플랫폼에서는 상품 카테고리에 따른 분포 차이로 인해 학습된 모델의 예측 성능 저하로 이어질 수 있다. 본 연구는 리뷰 유용성 예측 문제를 새로운 상품 카테고리로의 일반화 문제로 확장하고 도메인 일반화 알고리즘의 적용 가능성과 한계를 실증적으로 분석하였다. 이를 위해 Amazon Reviews 2023 데이터셋의 33개 상품 카테고리를 각각 하나의 도메인으로 정의하였다. 이후, 15개 도메인을 고정된 소스 도메인으로 설정하고 통합하여 학습 데이터로 사용하고, 이를 학습에 사용되지 않은 나머지 도메인으로 평가하여 고정 소스-다중 타겟 설정에서 미학습 타겟 도메인에 대한 성능을 측정하였다. 도메인 간 차이를 고려하지 않는 ERM을 베이스라인으로 설정하여 대표적인 도메인 일반화 알고리즘인 DeepCORAL, DANN, GroupDRO, MLDG, Mixup과 동일한 조건에서 비교하였다. 실험 결과를 통해 리뷰 유용성 예측에서 도메인 일반화 알고리즘의 적용이 항상 성능 향상으로 이어지는 것은 아니며, 알고리즘이 도메인 정보를 활용하고 조정하는 방식이 미학습 타겟 도메인 성능에 영향을 미칠 가능성을 시사하였다.

[Abstract]

This study extends review helpfulness prediction to the problem of generalization across unseen product categories and empirically evaluates domain generalization algorithms. To this end, 33 product categories in the Amazon Reviews 2023 dataset were each treated as a separate domain. Among these, 15 domains were fixed as source domains and combined to construct the training data, while the remaining 18 domains were reserved for evaluation. This fixed-source, multiple-target setting was employed to measure performance on unseen domains. Empirical Risk Minimization (ERM), which does not explicitly account for domain differences, was used as the baseline. Under identical experimental conditions, ERM was compared with representative domain generalization algorithms, including DeepCORAL, DANN, GroupDRO, MLDG, and Mixup. The results show that domain generalization algorithms do not necessarily improve review helpfulness prediction performance and suggest that how these algorithms utilize and adjust domain-specific information can significantly affect generalization to unseen domains.

색인어 : 리뷰 유용성 예측, 도메인 일반화, 온라인 리뷰, 도메인 이동, BERT
Keyword : Review Helpfulness Prediction, Domain Generalization, Online Reviews, Domain Shift, BERT

http://dx.doi.org/10.9728/dcs.2026.27.10.2835

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-CommercialLicense(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 30 June 2026; **Revised** 03 August 2026
Accepted 31 August 2026

***Corresponding Author; Jeongeun Park**

Tel: +82-31-400-1058
E-mail: parkje@hanyang.ac.kr

Ⅰ. 서 론

온라인 리뷰는 소비자의 구매 의사결정을 지원하는 핵심 정보 자원으로 활용되고 있으며[1] 전자상거래에서 매우 보편화되어 있다[2]. 그러나, 방대한 양과 일관되지 않은 품질로 정보 과부하를 야기하기도 한다[3]. 이를 해결하기 위해 상당수의 온라인 리뷰 플랫폼이 유용한 리뷰를 식별하는 데 도움이 되도록 유용성 투표 시스템을 구축하였다. 유용성 투표 시스템은 사용자가 유용하다고 판단한 리뷰에 투표하도록 하고, 사용자가 추천한 리뷰를 우선 노출하여 방대한 리뷰 속에서 양질의 정보를 빠르게 식별할 수 있게 한다. 그러나, 신규 리뷰나 비인기 제품 리뷰의 경우, 유용성 투표를 받을 기회 자체가 부족하다는 한계가 있다[4]-[6]. 이에 따라, 리뷰 텍스트, 평점, 사용자 정보 등의 메타 데이터를 활용하여 리뷰의 유용성을 자동으로 예측하는 연구가 지속적으로 수행되어 왔다.

기존 리뷰 유용성 예측 연구는 대부분 i.i.d.(independent and identically distributed) 가정을 기반으로 수행되어 왔다. i.i.d. 가정이란 학습 데이터와 테스트 데이터가 동일한 분포를 따른다는 가정이다[7]. i.i.d. 가정 하에 학습된 모델은 학습 데이터의 분포와 다른 분포의 데이터가 입력되는 o.o.d.(out-of-distribution) 상황에서 성능이 크게 저하되는 경향이 있다[8]. 이처럼 학습 데이터와 테스트 데이터 간의 분포 차이로 모델 성능이 하락하는 현상을 도메인 이동이라고 한다[9],[10]. 모델 학습 시에 항상 테스트 데이터에 접근 가능한 것이 아니므로 도메인 이동은 딥러닝 모델 일반화 측면에서 중요한 문제로 다뤄진다[10]. 즉, 기존 연구들은 주로 동일 분포 기반 환경에서 모델 성능을 평가하였으므로 도메인 이동 환경에서의 일반화 성능은 충분히 고려하지 않았다는 것이다.

반면, 온라인 리뷰는 전자상거래, 숙박, 영화, 음악, 비디오 콘텐츠 등 다양한 디지털 플랫폼을 넘나들며 활용되기 때문에[11] 실제 서비스 환경에서는 리뷰 길이, 표현 방식, 평점 분포 등 데이터 분포가 플랫폼과 상품 카테고리에 따라 크게 달라질 수 있다. 예를 들어, 전자제품 리뷰는 성능 및 기능 중심으로 리뷰가 작성될 수 있고 뷰티 제품의 경우 개인 경험 중심으로 작성될 수 있다. 본 연구에서 사용한 Amazon 리뷰 데이터의 기술통계 결과(표 1), 실제 리뷰 데이터에서 상품 카테고리별로 리뷰 길이, 유용성 투표 수, 상품별 누적 평점 수(Rating Number)의 분포에 차이가 나타남을 확인할 수 있었다. 상품 카테고리에 따른 리뷰 차이는 학습 데이터와 실제 예측되는 데이터 간의 분포 차이를 유발하고 기존 i.i.d. 기반 모델의 일반화 성능을 제한한다. 따라서, o.o.d. 상황이 빈번한 실제 리뷰 유용성 예측 환경에서도 안정적인 성능을 유지할 수 있는 모델이 필요하다.

이러한 도메인 이동 문제를 해결하기 위한 접근 방식은 크게 도메인 적응과 도메인 일반화, 두 가지로 구분된다[12]. 도메인 이동에서 데이터는 학습 사용 여부에 따라 소스 데이터와 타겟 데이터로 나뉜다. 소스 데이터는 모델 학습에 사용되는 데이터이고 타겟 데이터는 학습에 사용되지 않았지만 실제 예측 대상이 되는 데이터를 의미한다. 도메인 적응은 전이학습의 일종으로, 타겟 도메인의 데이터를 활용하여 이미 학습된 모델을 새로운 환경에 빠르게 적응시키는 것을 목표로 하는 반면[12],[13], 도메인 일반화는 타겟 도메인에 대한 정보 없이, 학습된 모델이 새로운 도메인에서도 안정적으로 동작하는 것을 목표로 한다[7],[13]. 도메인 일반화는 타겟 도메인에 접근할 수 없고 실제 예측 데이터에 맞춰 재학습할 수 없다는 점에서 도메인 적응보다 해결 난도가 높지만, 실제 서비스를 고려한 현실적인 접근 방식으로 평가받는다[7],[8],[10].

따라서, 리뷰 유용성 예측의 현실적인 배포 환경을 고려하여, 타겟 도메인에 대한 사전 정보 없이 안정적인 예측 성능을 확보할 수 있는 도메인 일반화 관점에서의 분석이 필요한 시점이다. 본 연구는 기본 학습 방식인 ERM(Empirical Risk Minimization)을 베이스라인으로 설정하고, 도메인 일반화의 대표적인 세 가지 접근 방식인 데이터 조작, 표현 학습, 학습 전략 기반에 해당하는 알고리즘을 비교한다. 이를 통해 도메인 일반화 알고리즘을 사용하는 것이 기본 학습 방식 대비 실제 성능 향상으로 이어지는지 검증하고, 알고리즘별 도메인 정보 활용 방식이 리뷰 유용성 예측에 미치는 영향을 확인한다. 이러한 결과를 바탕으로 본 연구는 리뷰 유용성 예측 문제를 새로운 상품 카테고리로의 일반화 문제로 확장하고, 실제 서비스 환경에서 도메인 일반화 알고리즘을 적용할 때 고려해야 할 실증적 근거를 제공한다.

Ⅱ. 선행연구

2-1 리뷰 유용성 예측

온라인 리뷰 유용성 예측 초기 연구는 선형 회귀, 랜덤 포레스트, 서포트 벡터 머신 등 전통적인 머신러닝 기법을 중심으로 수행되었다. 이러한 머신러닝 기반 접근은 리뷰 텍스트나 메타 데이터로부터 유용성 판단에 필요한 특징을 수작업으로 추출하는 방식에 의존하였기 때문에, 추출된 특징 표현의 한계가 예측 성능의 제약으로 이어졌다[14],[15].

이러한 한계를 보완하기 위해 최근 신경망 기반 모델을 활용하여 리뷰의 특징 표현을 자동으로 학습하려는 연구가 진행되었다. Olmedilla et al.[16]은 1D-CNN을 활용하여 온라인 리뷰 유용성을 예측하였고 Kong et al.[17]은 CNN 기반 주 모델과 수작업 특징 기반 보조 모델을 결합한 하이브리드 구조를 제안하였다. 그들은 CNN을 통해 리뷰 텍스트의 의미 정보를 학습하고, TransE를 활용하여 리뷰, 리뷰어, 제품 간 관계 정보를 임베딩함으로써 리뷰 유용성 예측 성능을

표 1. 상품 카테고리별 리뷰 데이터의 기술통계
Table 1. Descriptive statistics of review data by product category

Category	Reviews(N)	Review Length Median[IQR]	Helpful Vote Median[IQR]	Products(N)	Rating Number Median[IQR]
All_Beauty	181,349	33.0 [16.0-64.0]	1.0 [1.0-3.0]	47,909	21.0 [7.0-60.0]
Amazon_Fashion	475,849	31.0 [15.0-59.0]	1.0 [1.0-2.0]	240,817	7.0 [3.0-18.0]
Appliances	395,404	43.0 [20.0-85.0]	1.0 [1.0-3.0]	42,292	46.0 [12.0-162.0]
Arts_Crafts_and_Sewing	2,209,939	32.0 [16.0-60.0]	1.0 [1.0-3.0]	326,235	42.0 [13.0-135.0]
Automotive	3,982,999	37.0 [18.0-72.0]	1.0 [1.0-3.0]	741,272	27.0 [9.0-83.0]
Baby_Products	1,129,829	52.0 [26.0-95.0]	1.0 [1.0-3.0]	111,591	49.0 [13.0-197.0]
Beauty_and_Personal_Care	6,061,950	37.0 [19.0-70.0]	1.0 [1.0-3.0]	554,941	48.0 [14.0-178.0]
Books	10,302,571	68.0 [31.0-148.0]	2.0 [1.0-4.0]	2,415,674	25.0 [7.0-116.0]
CDs_and_Vinyl	2,002,024	80.0 [35.0-167.0]	2.0 [1.0-5.0]	446,665	15.0 [5.0-59.0]
Cell_Phones_and_Accessories	2,958,574	42.0 [21.0-82.0]	1.0 [1.0-2.0]	447,966	42.0 [13.0-149.0]
Clothing_Shoes_and_Jewelry	12,685,494	35.0 [18.0-65.0]	1.0 [1.0-3.0]	2,609,391	29.0 [8.0-120.0]
Digital_Music	43,904	49.0 [22.0-114.0]	2.0 [1.0-3.0]	25,069	5.0 [2.0-22.0]
Electronics	9,090,843	55.0 [26.0-111.0]	1.0 [1.0-3.0]	816,584	31.0 [9.0-116.0]
Gift_Cards	20,713	24.0 [10.0-47.0]	1.0 [1.0-2.0]	791	278.0 [36.5-1211.5]
Grocery_and_Gourmet_Food	3,621,815	32.0 [16.0-60.0]	1.0 [1.0-3.0]	341,576	44.0 [12.0-174.0]
Handmade_Products	155,458	25.0 [13.0-47.0]	1.0 [1.0-2.0]	46,087	16.0 [6.0-46.0]
Health_and_Household	6,551,783	37.0 [18.0-71.0]	2.0 [1.0-3.0]	452,542	58.0 [16.0-226.0]
Health_and_Personal_Care	135,040	35.0 [17.0-69.0]	1.0 [1.0-3.0]	26,345	24.0 [8.0-70.0]
Home_and_Kitchen	14,939,452	39.0 [20.0-74.0]	1.0 [1.0-3.0]	1,660,385	36.0 [10.0-127.0]
Industrial_and_Scientific	1,146,728	38.0 [19.0-74.0]	1.0 [1.0-3.0]	165,203	32.0 [9.0-110.0]
Kindle_Store	6,248,119	56.0 [29.0-119.0]	1.0 [1.0-2.0]	901,144	73.0 [19.0-298.0]
Magazine_Subscriptions	30,555	37.0 [19.0-75.0]	2.0 [1.0-6.0]	2,522	30.0 [7.0-177.8]
Movies_and_TV	4,111,213	50.0 [23.0-114.0]	2.0 [1.0-4.0]	495,068	86.0 [17.0-513.0]
Musical_Instruments	818,842	50.0 [24.0-101.0]	2.0 [1.0-3.0]	116,762	20.0 [6.0-71.0]
Office_Products	2,546,275	40.0 [20.0-77.0]	1.0 [1.0-3.0]	305,802	43.0 [12.0-159.0]
Patio_Lawn_and_Garden	3,990,779	40.0 [20.0-76.0]	1.0 [1.0-3.0]	415,282	40.0 [12.0-131.0]
Pet_Supplies	3,632,505	47.0 [24.0-87.0]	1.0 [1.0-3.0]	270,239	46.0 [14.0-167.0]
Software	1,828,861	22.0 [11.0-41.0]	3.0 [1.0-9.0]	61,968	21.0 [7.0-86.0]
Sports_and_Outdoors	4,512,909	44.0 [21.0-85.0]	1.0 [1.0-3.0]	678,647	30.0 [9.0-109.0]
Subscription_Boxes	11,757	36.0 [17.0-68.0]	3.0 [2.0-8.0]	511	65.0 [20.0-198.0]
Tools_and_Home_Improvement	5,908,354	42.0 [21.0-81.0]	1.0 [1.0-3.0]	701,323	34.0 [10.0-122.0]
Toys_and_Games	3,384,305	38.0 [19.0-71.0]	1.0 [1.0-3.0]	442,343	56.0 [17.0-192.0]
Video_Games	1,124,995	60.0 [26.0-134.0]	2.0 [1.0-3.0]	85,200	42.0 [12.0-158.0]

향상시키고자 하였다. 연구[18]에서는 Bi-LSTM을 사용하여 리뷰 텍스트의 문맥적 표현을 생성하고, 어텐션 메커니즘을 통해 제품 메타데이터와 리뷰 텍스트 간의 관계를 학습하였다. Qu et al.[19]은 리뷰 감성과 제품 정보를 함께 고려하는 LSTM 기반 어텐션 모델을 제안하여, 고객 기대와 제품 정보가 리뷰 유용성 판단에 미치는 영향을 반영하고자 하였다. Bilal & Almazroi[3]는 Yelp 데이터셋을 활용하여 미세조정한 BERT 모델을 제안함으로써, 기존 수작업 특징에 대한 의존성을 낮추고자 하였다. Li et al.[20]은 BERT 기반 리뷰 유용성 예측 모델인 BHelPCorT를 제안하였고 이는 BERT 인코더와 어텐션 메커니즘을 활용하여 리뷰 텍스트와 별점 간 일관성 정보를 유용성 예측에 반영한 것이다.

대부분의 기존 리뷰 유용성 예측 연구는 학습 데이터와 테스트 데이터가 동일한 분포를 따른다는 i.i.d. 가정하에 수행되어 왔다. 그러나, 실제 서비스 환경에서는 학습에 사용된 도메인과 모델이 적용되는 테스트 도메인이 서로 다를 수 있으며, 이러한 cross-domain 상황에서의 리뷰 유용성 예측 연구는 상대적으로 부족한 실정이다. Chen et al.[21]은 도메인별로 충분한 라벨링 데이터가 필요하다는 기존 방법론의 한계와, 데이터가 부족하거나 out-of-vocabulary 문제가 발생하는 도메인에서 성능이 저하될 수 있다는 문제를 지적하였다. 이를 해결하기 위해 CNN 기반 리뷰 표현 학습 모듈과

보조 도메인 판별기를 결합한 cross-domain 리뷰 유용성 예측 모델을 제안하였다. Chen et al.[22]는 도메인에 따라 리뷰 유용성을 판단하는 데 중요한 정보가 달라질 수 있다는 점에 주목하여 Multi-Domain Gated CNN을 제안하였다. 이 모델은 단어, 문자 및 aspect 수준의 다중 표현에 domain-specific gating을 적용하여 도메인에 따라 유용성 판단에 중요한 정보를 차별적으로 반영하고, domain-specific output layer와 도메인 간 상관관계 학습을 통해 여러 도메인 간의 지식 전이를 수행한다. 이를 통해 데이터가 풍부한 소스 도메인의 정보를 활용하여 라벨링 데이터가 부족한 타겟 도메인의 리뷰 유용성 예측 성능을 향상시키고자 하였다. 표 2와 같이, [21],[22] 모두 소스 도메인에서 학습한 지식을 타겟 도메인으로 전이하는 transfer learning 관점으로, 타겟 도메인 데이터에 대한 사전 정보가 없는 도메인 일반화와는 큰 차이가 있다. 따라서, cross-domain 리뷰 유용성 예측에 대한 선행 연구는 존재하지만 타겟 도메인 정보를 활용하지 않는 도메인 일반화에서 리뷰 유용성 예측 문제를 다룬 연구는 아직 충분하지 않다. 본 연구는 이러한 연구 공백을 바탕으로 기본 학습 방식인 ERM과 대표적인 도메인 일반화 알고리즘 다섯 가지를 비교하여, 도메인 일반화 알고리즘의 적용이 미학습 타겟 도메인에서 성능 향상으로 이어지는지와 알고리즘별 도메인 정보 활용 방식이 성능에 미치는 영향을 분석한다.

표 2. 기존 Cross-domain 리뷰 유용성 예측 연구와 본 연구의 비교
Table 2. Comparison of prior cross-domain review helpfulness prediction studies and our study

Cross-domain RHP	Learning Paradigm	Training Data	Test Data	Test access
Chen et al. (2018)	Transfer Learning	Source & Target	Target	O
Chen et al. (2019)	Transfer Learning	Source & Target	Target	O
Ours	Domain Generalization	Source	Unseen Target	X

2-2 도메인 일반화

도메인 일반화는 훈련 데이터와 테스트 데이터 간에 분포 이동이 있을 경우, 성능이 저하되는 도메인 이동 문제를 해결하기 위한 방법이다[23]. 또한, 학습 과정에서 타겟 도메인 정보를 활용할 수 없는 상황을 해결하기 위한 접근법이기도 하다[24].

Wang et al.[7]은 도메인 일반화 알고리즘을 데이터 조작(data manipulation), 표현 학습(representation learning), 학습 전략(learning strategy) 기반 접근으로 분류하였다. 데이터 조작 기반 방법은 학습 데이터를 증강하거나[25] 생성하여 과적합을 방지함으로써 일반화 능력을 향상시킨다[23]. 한편, 표현 학습 기반 방식은 서로 다른 도메인에서도 일관되게 나타나는 특징 표현인 도메인 불변 표현을 추출하도록 하여 일반화 성능을 확보한다. 이는 적대적 학습[26], 도메인 특징 분포 정렬[27] 등의 방식으로 달성되었다. 학습 전략 기반 방법은 메타 학습[10],[28], 분포 강건 최적화[29] 등 모델을 훈련 시키는 방법의 변주를 통해 일반화 능력을 향상시킨다.

이와 같은 다양한 도메인 일반화 알고리즘을 동일한 조건에서 평가하려는 시도가 있었다. Gulrajani & Lopez[8]는 DomainBed 벤치마크를 통해 ERM, IRM, GroupDRO, Mixup, MLDG, CORAL, MMD, DANN 등을 Colored MNIST, Rotated MNIST, VLCS, PACS, Office-Home, Terra Incognita, DomainNet 데이터셋에서 비교하였다. 또한, Wang et al.[7]은 DeepDG 벤치마크를 제안하고 ERM, DANN, DeepCORAL, Mixup, RSC, GroupDRO, ANDMask 등을 PACS와 Office-Home 데이터셋에서 평가하였다. 두 연구는 공통적으로 도메인 이동 상황에서 ERM이 다양한 도메인 일반화 알고리즘과 비교해도 충분히 경쟁력 있는 성능을 보일 수 있음을 보고하였다. 특히, Wang et al.[7]은 데이터셋 내 도메인 간 분포 차이가 상대적으로 크지 않았기 때문에 이러한 결과가 도출되었을 것이라는 가능성을 제기하였다. 이는 도메인 일반화 알고리즘의 효과가 적용되는 데이터셋과 태스크의 특성에 따라 달라질 수 있음을 의미한다. 따라서, PACS, Office-Home, DomainNet과 같은 이미지 기반 데이터셋을 중심으로 수행된 기존 도메인 일반화 벤치마크 연구[7],[8]의 결과를 텍스트 기반의 리뷰 유용성 예측 태스크에 그대로 적용하기에는 한계가 있다. 이에, 본 연구는 리뷰 유용성 예측 문제에서 도메인 일반화 알고리즘의 적용이 미학습 타겟 도메인 성능 향상으로 이어지는지를 실증적으로 확인하고자 한다.

한편, 도메인 일반화는 학습에 사용되는 소스 도메인의 수에 따라, 단일 소스 도메인 일반화와 다중 소스 도메인 일반화로 구분될 수 있다. 단일 소스 도메인 일반화는 하나의 소스 도메인만을 활용하는 반면, 다중 소스 도메인 일반화는 여러 소스 도메인을 활용하여 미학습 타겟 도메인에서의 성능을 확보하는 것을 목표로 한다. 본 연구는 실제 온라인 플랫폼 환경에서 다양한 상품 카테고리로 일반화되는 상황을 고려하여 다중 소스 도메인 일반화 설정을 채택한다.

Ⅲ. 연구 방법

3-1 데이터셋 및 데이터 전처리

본 연구에서는 리뷰 유용성 예측 실험을 위해 Amazon Reviews 2023[30] 데이터셋을 사용하였다. Amazon Reviews 2023은 사용자가 작성한 리뷰 정보와 상품 관련 정

보인 메타 데이터가 분리되어 제공되므로, 두 데이터셋의 공통 변수이자 상품군을 식별하는 대표 상품 식별번호(Parent Asin)를 기준으로 병합하였다. 리뷰 데이터에서는 사용자가 작성한 실제 리뷰 본문(text), 리뷰 작성 시점(timestamp), 그리고 해당 리뷰가 다른 사용자로부터 받은 유용성 투표 수(helpful vote)를 활용하였다. 상품 메타 데이터에서는 상품의 누적 평점 수(rating number)를 활용하였고 이는 해당 상품에 대해 등록된 전체 평점 개수이다.

이와 같이 구성한 데이터를 바탕으로 리뷰 유용성 라벨을 정의하기에 앞서, 유용성 투표 수의 특성을 고려하여 분석 대상 리뷰를 선별하였다. 유용성 투표 수가 1 이상인 리뷰만을 사용하였다[18],[31]. 유용성 투표 수가 0인 리뷰는 실제로 유용하지 않은 리뷰뿐만 아니라, 최근에 작성되었거나 노출 기회가 적은 상품군에 속해 아직 투표를 받지 못한 리뷰일 가능성도 존재한다. 즉, 유용성 투표 수가 0인 리뷰를 일괄적으로 '유용하지 않음' 클래스로 포함하면 노출 부족으로 인해 투표를 받지 못한 양질의 리뷰를 '유용하지 않음' 클래스로 오분류할 가능성이 있다. 또한, Amazon 리뷰 데이터에서는 유용성 투표를 받지 못한 리뷰가 다수 포함되어 있으므로, 이를 포함한 상태에서 라벨링 기준을 산정할 경우 '유용함' 클래스의 기준이 과도하게 낮아질 수 있다. 이에, 본 연구에서는 투표 수가 0인 리뷰를 제외하였다.

리뷰 유용성 예측을 이진 분류 문제로 정의하고 유용성 투표 수를 기준으로 라벨링을 진행하였다. 기존 연구에서는 '유용함' 투표 수를 전체 투표 수로 나눈 유용성 비율을 라벨링 기준으로 주로 사용하였다. 여기서 전체 투표 수는 '유용함' 투표 수와 '유용하지 않음' 투표 수를 더한 값이다. 그러나, 최근 아마존이 '유용하지 않음' 투표 버튼을 삭제하면서 기존의 유용성 비율을 라벨링 기준으로 사용하는 것이 불가능해졌다. 이에, 본 연구에서는 플랫폼의 변화를 반영하여 '유용함' 투표 개수를 기반으로 라벨을 정의하였다. Amazon 데이터셋의 각 카테고리를 하나의 도메인으로 정의하였고 각 도메인 내부에서 유용성 투표 수의 상위 25%에 해당하는 리뷰를 '유용함' 클래스로, 그 미만인 리뷰를 '유용하지 않음' 클래스로 정의하였다. 따라서, 본 연구에서의 리뷰 유용성은 모든 도메인에 공통으로 적용되는 절대적인 투표 수가 아니라, 동일한 도메인 내에서 상대적으로 높은 유용성 투표를 획득한 정도를 의미한다. 이는 상품 도메인마다 리뷰 수와 유용성 투표 수 분포가 다르므로, 전체 데이터에 동일한 기준을 적용할 경우 특정 도메인의 리뷰가 '유용함' 또는 '유용하지 않음' 클래스로 과도하게 분류될 수 있기 때문이다. 각 도메인에서 '유용함' 클래스로 분류되기 위한 최소 투표수 기준, 즉, 유용성 투표수의 상위 25% 지점인 제3사분위수(Quartile 3, Q3)는 표 3과 같다. 표 3에서 도메인별 제3사분위수가 2에서 9까지 다르게 나타나는 것을 통해 도메인별로 이용자 참여 수준 및 투표 분포의 차이가 있음을 알 수 있다. 각 도메인별 유용성 투표 수의 제3사분위수와 리뷰별 유용성 투표 수는 정답 라벨을 구성하기 위한 용도로만 사용하였으며, 모델의 학습 입력이나 테스트 입력에는 포함하지 않았다. 따라서 도메인별 제3사분위수는 모델 학습이나 추론에 활용되는 정보가 아니라, 각 도메인 내 상대적 유용성을 정의하고 예측 성능을 평가하기 위한 ground truth 구성 기준으로만 사용된다.

표 3. 도메인별 라벨링 기준
Table 3. Domain-specific labeling criteria

Category	Q3	Category	Q3
All_Beauty	3	Health_and_Personal_Care	3
Amazon_Fashion	2	Home_and_Kitchen	3
Appliances	3	Industrial_and_Scientific	3
Arts_Crafts_and_Sewing	3	Kindle_Store	2
Automotive	3	Magazine_Subscriptions	6
Baby_Products	3	Movies_and_TV	4
Beauty_and_Personal_Care	3	Musical_Instruments	3
Books	4	Office_Products	3
CDs_and_Vinyl	5	Patio_Lawn_and_Garden	3
Cell_Phones_and_Accessories	2	Pet_Supplies	3
Clothing_Shoes_and_Jewelry	3	Software	9
Digital_Music	3	Sports_and_Outdoors	3
Electronics	3	Subscription_Boxes	8
Gift_Cards	2	Tools_and_Home_Improvement	3
Grocery_and_Gourmet_Food	3	Toys_and_Games	3
Handmade_Products	2	Video_Games	3
Health_and_Household	3		

텍스트 전처리 과정에서는 문자열 양 끝 공백 제거, 소문자 변환, URL 및 이메일 주소 제거, 이모티콘 제거, 연속 공백 축소 등의 작업을 수행하였다. 또한 리뷰 작성 시점(timestamp) 정보를 활용하여 연(year), 월(month), 일(day) 정보를 추가 변수로 생성하였다.

3-2 도메인 일반화 알고리즘

도메인 일반화 알고리즘은 접근 방식에 따라 데이터 조작, 표현 학습, 학습 전략의 세 가지 유형으로 분류할 수 있다[7]. 본 연구에서는 도메인 일반화의 다양한 접근 방식별 성능 차이를 체계적으로 비교하기 위해 각 유형을 대표하는 알고리즘을 고루 선정하였다. 표현 학습 기반으로 DANN(Domain-Adversarial Neural Network)과 DeepCORAL(Deep neural networks Correlation Alignment), 학습 전략 기반으로 MLDG(Meta-Learning for Domain Generalization)와 GroupDRO(group Distributionally Robust Optimization), 데이터 조작 기반으로 Mixup을 사용하였다. 또한, 가장 기본적인 학습 방식으

로써 ERM을 베이스라인으로 설정하였다.

1) ERM

ERM[32]는 제한된 학습 데이터에서 계산된 평균 손실, 즉 경험적 위험을 최소화하도록 모델을 학습하는 방식이다. 일반적으로 모델 학습의 이상적인 목표는 실제 데이터 분포에서의 예측 오차를 최소화하는 것이지만, 실제 데이터 분포를 완벽하게 알 수 없다는 한계가 있다. 따라서, ERM은 관측 가능한 학습 데이터에서의 손실을 기준으로 모델을 최적화한다. 이를 위해 각 학습 샘플에 대한 예측값과 실제 라벨 간의 손실을 계산하고 이 값들의 평균인 경험적 위험을 최소화하도록 모델의 파라미터를 업데이트한다. ERM의 목적 함수는 식 (1)과 같다.

$$\arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(f_\theta(x_i), y_i) \tag{1}$$

N은 전체 학습 샘플 수, x_i는 i번째 입력, y_i는 해당 입력의 라벨을 의미한다. $f_\theta(x_i)$는 파라미터 θ를 갖는 모델이 x_i에 대해 출력한 예측값이며, $l(f_\theta(x_i), y_i)$는 예측값과 실제 라벨 간의 손실이다. 도메인 일반화 알고리즘이 타겟 도메인에서의 성능 하락을 방지하기 위해 학습을 수행하는 것과 달리, ERM은 도메인 간 차이를 고려하지 않고, 여러 소스 도메인의 데이터를 하나의 통합된 학습 데이터로 간주하여 분류 손실을 최소화한다. 따라서, 본 연구에서는 ERM을 도메인 일반화 알고리즘의 성능 향상 여부를 판단하기 위한 베이스라인으로 설정하였다.

2) 표현 학습 기반

DANN[26]은 적대적 학습 기반 알고리즘이다. 모델은 특징 추출기 F, 클래스 분류기 C, 도메인 분류기 D로 구성된다. DANN의 핵심 아이디어는 특징 추출기가 클래스 분류에는 유용하지만 도메인 자체는 구분하기 어려운 도메인 불변 표현을 학습하도록 하는 것이다. 최종적인 DANN의 목적함수는 식 (2)와 같다.

$$\min_{\theta_F, \theta_C} \max_{\theta_D} L_{class} - \lambda_{adv} L_{domain} \tag{2}$$

여기서 θ_F, θ_C, θ_D는 각각 특징 추출기, 클래스 분류기, 도메인 분류기의 파라미터이고, λ_{adv}는 도메인 적대적 손실의 반영 비중을 조절하는 하이퍼 파라미터이다. 목적함수의 최적화를 위한 과정에서 GRL(gradient reversal layer)이 사용된다. GRL은 순전파 과정에서는 입력을 그대로 통과시키지만, 역전파 과정에서는 도메인 분류 손실로부터 들어오는 그라디언트의 방향을 반대로 역전시켜 특징 추출기에 전달함으로써, 특징 추출기가 클래스 분류 손실은 최소화하면서 동시에 도메인 분류기가 도메인을 구분하지 못하도록 학습시킨다.

DANN은 레이블이 존재하는 소스 도메인과 레이블이 없는 타겟 도메인 간의 분포 차이를 완화하기 위해 제안된 도메인 적응 알고리즘이다. 본 연구에서는 타겟 도메인 데이터를 학습에 사용하지 않는 다중 소스 도메인 일반화 설정에 맞게 DANN을 적용하였다. 학습에 사용되는 여러 상품 도메인을 각각 하나의 소스 도메인으로 정의하고 도메인 분류기는 입력 리뷰의 표현이 여러 소스 도메인 중 어느 도메인에서 생성되었는지를 예측하는 다중 클래스 분류기로 학습되었다. 반면 특성 추출기는 클래스 분류기의 리뷰 유용성 예측 손실을 최소화하는 동시에, GRL을 통해 전달되는 도메인 분류 손실의 역방향 기울기를 이용하여 소스 도메인 간 구분이 어려운 표현을 학습하였다. 즉, 본 연구에서 타겟 도메인의 리뷰, 유용성 라벨 및 도메인 분포 정보는 모델 학습에 사용하지 않았고, DANN은 타겟 도메인과 소스 도메인을 정렬하는 방식이 아니라, 여러 소스 도메인 사이에서 공통적인 도메인 불변 표현을 학습하도록 구현하였다.

DeepCORAL[27]은 도메인 간 특징 분포를 통계적으로 정렬하기 위한 방법이다. DeepCORAL의 핵심 아이디어는 서로 다른 도메인의 특징 공분산 행렬을 정렬함으로써 도메인 간 분포 차이를 감소시키는 것이다. CORAL 손실은 식(3)와 같이 정의된다.

$$L_{CORAL} = \frac{1}{4d^2} \| C_S - C_T \|_F^2 \tag{3}$$

식 (3)에서 C_S와 C_T는 각각 소스 도메인과 타겟 도메인의 특징 공분산 행렬을 의미한다. d는 특징 차원 수이며, $\| \cdot \|_F^2$는 두 행렬 간의 거리를 측정하기 위한 Frobenius Norm의 제곱이다. 최종 목적 함수는 식 (4)와 같다.

$$\arg\min_{\theta} L_{CLASS} + \lambda L_{CORAL} \tag{4}$$

즉, 모델 파라미터를 학습시키기 위한 최종 손실은 메인 태스크를 위한 클래스 분류 손실 L_{CLASS}과 CORAL 손실의 가중합으로 구성된다. 여기서 λ는 CORAL 손실의 반영 비중을 조절하는 하이퍼파라미터이다.

본 연구에서는 도메인 일반화 설정을 따르고 있으므로, 타겟 도메인의 특징 공분산을 학습에 사용하지 않았다. 구체적으로, 각 미니배치에서 두 개 이상의 샘플을 포함하는 각 소스 도메인의 특징 공분산 행렬을 계산하였다. 이후 미니배치에 포함된 서로 다른 모든 소스 도메인 쌍에 대해 CORAL 손실을 계산하고, 이를 평균하여 최종 CORAL 손실로 사용하였다. 하나의 미니배치에 공분산을 계산할 수 있는 소스 도메인이 두 개 미만인 경우에는 CORAL 손실을 적용하지 않았다. 따라서 본 연구에서 DeepCORAL의 정렬 대상은 소스 도메인과 타겟 도메인이 아니라, 학습에 사용된 소스 상품 도메인

들 간의 특징 분포이다.

3) 학습 전략 기반

MLDG[10]는 도메인 일반화를 위한 대표적인 메타러닝 기반 알고리즘이다. MLDG는 학습 과정에서 도메인 이동을 반복적으로 시뮬레이션함으로써, 타겟 도메인에서의 성능 하락을 최소화하는 것을 목표로 한다. 학습 과정에서 도메인 이동을 시뮬레이션하기 위해 소스 도메인을 meta-train 도메인과 meta-test 도메인으로 분할하고 meta-train 도메인에서 업데이트된 파라미터를 사용하여 meta-test 도메인에서의 손실을 계산한다. meta-train 손실은 식 (5), meta-test 손실은 식 (6)과 같다.

$$L_{meta-train} = S(\theta) \tag{5}$$

$$L_{meta-test} = T(\theta - \alpha S'(\theta)) \tag{6}$$

식 (6)에서 α는 가상 업데이트를 위한 inner learning rate을 의미한다. MLDG의 최종 목적 함수는 식 (7)과 같이 meta-train 손실과 meta-test 손실의 가중합으로 정의되고 β는 meta-test 손실의 반영 비중을 조절하는 하이퍼파라미터이다.

$$\arg\min_{\theta} L_{meta-train} + \beta L_{meta-test} \tag{7}$$

GroupDRO[29]는 학습 과정에서 계산된 도메인별 손실이 가장 큰 최악의 도메인(worst-case group)에서의 성능 저하를 완화하도록 학습하는 알고리즘이다. 먼저, 도메인 g에 대한 평균 손실은 식 (8)과 같이 정의된다.

$$L_g(\theta) = \frac{1}{N_g} \sum_{i:d_i=g} l(f_\theta(x_i), y_i) \tag{8}$$

여기서 N_g는 도메인 g에 속한 학습 샘플 수이며, $l(f_\theta(x_i), g_i)$는 모델의 예측값과 실제 라벨 간의 손실이다. $d_i = g$를 통해 해당 도메인에 속하는 데이터의 손실만 취한다. GroupDRO의 목적함수는 식 (9)와 같이 정의된다.

$$\min_{\theta} \max_{g \in \Delta_G} \sum_{g=1}^{G} q_g L_g(\theta) \tag{9}$$

식 (9)에서 G는 전체 소스 도메인의 수, q_g는 도메인 g에 부여되는 가중치를 의미한다. Δ_G는 G개 도메인에 대한 가중치 집합으로, 각 가중치는 0 이상의 값을 가지고 전체 합은 1이 되도록 제한된다. 즉, GroupDRO는 손실이 큰 도메인에 더 큰 가중치를 부여하여, 평균 손실만을 최소화하는 ERM과 달리 성능이 낮은 도메인의 손실을 줄이는 데 초점을 둔다.

4) 데이터 조작 기반

Mixup[25]는 두 개의 학습 샘플과 레이블을 선형 보간하여 새로운 학습 샘플을 생성하는 데이터 증강 기반 기법이다. Mixup은 모델이 학습 샘플 사이의 영역에서도 선형적으로 동작하도록 함으로써 보다 부드러운 결정 경계를 학습하도록 유도한다. 이를 통해 과적합을 완화하고 일반화 성능 향상에 기여한다. 가상 샘플을 생성하기 위한 학습 샘플 및 라벨 변환 식은 각각 식 (10)과 식 (11)과 같다.

$$\tilde{x} = \lambda x_i + (1-\lambda) x_j \tag{10}$$

$$\tilde{y} = \lambda y_i + (1-\lambda) y_j \tag{11}$$

여기서 x_i, x_j는 미니배치 내에서 랜덤하게 선택된 두 개의 입력 데이터이며, y_i, y_j는 각각에 대응하는 실제 라벨이다. $\tilde{x}$와 $\tilde{y}$는 선형 결합에 의해 생성된 가상의 혼합 샘플과 혼합 라벨이다. 두 데이터가 혼합되는 비중을 λ로 조정할 수 있는데 이때 λ는 일반적으로 Beta 분포로부터 샘플링된다. λ가 0 또는 1에 가까울수록 원본 샘플이 생성되며 λ가 0.5일 때 가장 균형적으로 혼합된 가상 샘플이 생성된다. 최종적으로 Mixup 모델을 최적화하기 위한 목적 함수는 식 (12)와 같다.

$$\arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(f_\theta(\tilde{x}), \tilde{y}) \tag{12}$$

식 (12)에서 N은 총 샘플 수이고, $l(f_\theta(\tilde{x}), \tilde{y})$는 가상 입력에 대한 모델의 예측값과 혼합 라벨 사이의 개별 손실 함수를 나타낸다. Mixup은 이 목적 함수를 최소화함으로써 모델이 학습 샘플 사이의 영역에서도 선형적으로 동작하도록 유도하고, 궁극적으로 보다 부드러운 결정 경계를 형성하게 한다. 이를 통해 특정 도메인에 대한 과적합을 완화하고, 타겟 도메인에 대한 일반화 성능 향상에 기여한다.

3-3 평가 프로토콜

타겟 도메인 분포에 접근할 수 없다는 도메인 일반화의 가정에 따라, 도메인 일반화 성능 평가를 위해 학습에 한 번도 사용되지 않은 미학습 타겟 도메인에 대한 모델의 성능을 평가하는 방식으로 LODO(Leave-One-Domain_Out)와 FSMT(Fixed Source Multiple Target)가 있다.

LODO는 전체 도메인 중 하나를 타겟 도메인으로 분리하고, 나머지 도메인을 이용하여 모델을 학습하는 방식이다. 학습된 모델은 사전에 분리된 타겟 도메인에서 평가되며, 이러한 과정을 각 도메인이 한 번씩 타겟 도메인으로 사용되도록 반복 수행한다. 즉, 도메인 수만큼 서로 다른 모델이 학습되고 미학습 타겟 도메인으로써의 역할을 수행하는 각 타겟 도메인에 대한 성능을 평가한 뒤 평균 성능을 최종 결과로 보고한다[8].

반면, FSMT는 학습에 사용되는 소스 도메인 집합을 고정한 뒤, 학습에 사용되지 않은 여러 타겟 도메인에 대해 모델의 일반화 성능을 평가하는 방식이다[33]. LODO는 전체 도메인 수만큼 반복적으로 모델을 학습해야 하는 반면, FSMT는 고정된 소스 도메인에서 모델을 한 번만 학습한 후 이를 여러 개의 미학습 타겟 도메인에 적용한다. 이러한 설정은 별도의 재학습 없이 적용해야 하는 실제 예측 환경을 보다 현실적으로 반영한다. 따라서, 본 연구에서는 현실 적용 환경과 유사한 FSMT 프로토콜을 기반으로 도메인 일반화 성능을 평가한다.

Ⅳ. 실 험

Amazon 리뷰 데이터셋의 각 카테고리를 하나의 도메인으로 정의하고, 데이터셋에서 사용 가능한 총 33개의 도메인을 대상으로 실험을 수행하였다. 앞서 서술한 데이터 전처리 과정에 따라, 유용성 투표 수가 1 이상인 리뷰만을 남긴 후, 텍스트 전처리 과정을 수행한 이후 각 도메인 내 유용성 투표 수의 제3사분위수 이상인 리뷰를 '유용함' 클래스로, Q3 미만인 리뷰를 '유용하지 않음' 클래스로 정의하였다. 도메인별 데이터 규모 차이와 클래스 불균형에 의한 문제를 배제하기 위해 각 도메인에서 '유용함' 리뷰 1,500개와 '유용하지 않음' 리뷰 1,500개를 랜덤 샘플링하여 도메인별 3,000개의 리뷰로 총 99,000개의 리뷰 데이터를 구성하였다. 이후 각 도메인별로 학습, 검증, 테스트 세트로 각각 9:0.5:0.5의 비율로 분할하였다. 각 도메인의 원자료 규모와 샘플링 이전의 클래스별 리뷰 개수는 표 4에 제시하였다.

FSMT 평가 프로토콜에 따라, 전체 33개 도메인 중 학습

표 4. 상품 도메인별 원자료 규모 및 균형 샘플링 이전 클래스 분포
Table 4. Raw review counts and class distribution before balanced sampling by domain

Category	Raw Reviews (N)	Candidate Pool(N)	Candidate Unhelpful(N)	Candidate Helpful(N)
All_Beauty	701,528	181,349	128,068	53,281
Amazon_Fashion	2,500,939	475,849	275,070	200,779
Appliances	2,128,605	395,404	271,857	123,547
Arts_Crafts_and_Sewing	8,966,758	2,209,939	1,529,764	680,175
Automotive	19,955,450	3,982,999	2,960,711	1,022,288
Baby_Products	6,028,884	1,129,829	811,944	317,885
Beauty_and_Personal_Care	23,911,390	6,061,950	4,220,028	1,841,922
Books	29,475,453	10,302,571	7,403,757	2,898,814
CDs_and_Vinyl	4,827,273	2,002,024	1,492,705	509,319
Cell_Phones_and_Accessories	20,812,945	2,958,574	1,751,852	1,206,722
Clothing_Shoes_and_Jewelry	66,033,346	12,685,494	9,179,401	3,506,093
Digital_Music	130,434	43,904	29,890	14,014
Electronics	43,886,944	9,090,843	6,316,512	2,774,331
Gift_Cards	152,410	20,713	12,271	8,442
Grocery_and_Gourmet_Food	14,318,520	3,621,815	2,595,670	1,026,145
Handmade_Products	664,162	155,458	89,652	65,806
Health_and_Household	25,631,345	6,551,783	4,427,146	2,124,637
Health_and_Personal_Care	494,121	135,040	92,760	42,280
Home_and_Kitchen	67,409,944	14,939,452	10,571,312	4,368,140
Industrial_and_Scientific	5,183,005	1,146,728	809,166	337,562
Kindle_Store	25,577,616	6,248,119	3,769,397	2,478,722
Magazine_Subscriptions	71,497	30,555	22,089	8,466
Movies_and_TV	17,328,314	4,111,213	3,054,257	1,056,956
Musical_Instruments	3,017,439	818,842	548,441	270,401
Office_Products	12,845,712	2,546,275	1,825,661	720,614
Patio_Lawn_and_Garden	16,490,047	3,990,779	2,735,241	1,255,538
Pet_Supplies	16,827,862	3,632,505	2,606,516	1,025,989
Software	4,880,181	1,828,861	1,365,154	463,707
Sports_and_Outdoors	19,595,170	4,512,909	3,153,206	1,359,703
Subscription_Boxes	16,216	11,757	8,717	3,040
Tools_and_Home_Improvement	26,982,256	5,908,354	4,147,779	1,760,575
Toys_and_Games	16,260,406	3,384,305	2,399,509	984,796
Video_Games	4,624,615	1,124,995	747,933	377,062

에 사용할 소스 도메인 K개를 선정하고 선택되지 않은 도메인을 미학습 타겟 도메인으로 사용하였다. 다수의 상품 도메인을 학습에 포함하여 소스 도메인의 다양성을 확보하는 동시에, 충분한 수의 미학습 타겟 도메인을 평가 대상으로 유지하기 위하여 전체 도메인 33개의 절반 규모인 15로 K를 고정하고 나머지 18개를 모델 평가에 사용하였다. 이후 선정된 소스 도메인들의 학습, 검증, 테스트 데이터를 각각 통합하여 최종 학습 세트, 검증 세트, 테스트 세트를 구성하였다. 학습에 사용될 소스 도메인은 랜덤하게 선정하였다. 또한, 도메인을 랜덤으로 선택함에 따라 특정 소스 도메인 조합에 결과가 종속되는 것을 방지하고, 소스-타겟 도메인 구성 변화에 따른 성능을 확인하기 위해 서로 다른 세 개의 도메인 선택 seed를 사용하여 세 가지 FSMT 분할을 구성하였다. 각 seed 별 분할 성능은 해당 분할에 포함된 18개 미학습 타겟 도메인 성능의 Macro F1-score 평균으로 계산하였다. 이후 세 분할의 평균을 종합하여 보고하였다. random seed에 따라 선택된 소스 도메인은 표 5에, 이에 따라 결정된 미학습 타겟 도메인은 표 6에 제시하였다.

본 연구에서는 리뷰 텍스트와 수치형 메타 데이터를 함께 활용하여 리뷰 유용성을 예측하였다. 텍스트 입력은 사전학습 언어모델인 BERT[34]를 사용하여 인코딩하였고 최종 은닉 상태에서 [CLS] 토큰을 문장 수준 텍스트 표현으로 사용하였다. 수치형 특성으로 연(year), 월(month), 일(day), 상품 누적 평점 수(rating number)를 사용하였고 해당 특성들은 linear layer와 ReLU 활성화 함수로 구성된 projection layer를 거쳐 128차원의 은닉 표현으로 변환되었다. 이후 BERT 기반 텍스트 표현과 projection layer를 통과한 수치형 특성 표현을 결합하였다. 결합된 특징은 fully connected layer와 ReLU 활성화 함수를 거쳐 최종 특징 표현으로 변환되었고 마지막으로 선형 분류기를 통해 리뷰 유용성 여부를 분류하였다. 본 연구에서는 Bert-base-uncased 사전학습 모델과 이에 대응하는 AutoTokenizer를 사용하였다. 모델 학습 시 BERT 입력 시퀀스의 최대 길이는 320으로 설정하였고 batch size는 32, 학습률은 2e-5, 그리고 epoch 수는 10으로 설정하였다. 또한, 모델 최적화에는 AdamW optimizer를 사용하였고 weight decay는 0.01로 설정하였으며, 학습 효율성을 높이기 위해 mixed precision을 적용하였다. 최종 모델은 10 epoch 동안 학습을 수행한 후 검증 세트에서의 Macro F1-score가 가장 높은 시점의 모델을 저장하였다. 모든 실험에서 Python의 random, NumPy 및 PyTorch의 CPU·CUDA에 동일한 seed 42를 적용하였다. 실험은 NVIDIA GeForce RTX 2080 Ti GPU 환경에서 수행되었으며, Python 3.10.20, PyTorch 2.9.1, CUDA 13.0, Transformers 4.57.3 및 scikit-learn 1.7.2를 사용하였다.

DANN에서는 리뷰 유용성 분류를 위한 클래스 분류 손실과 도메인 분류를 위한 도메인 분류 손실 사이의 상대적 비중을 동일 비중으로 반영하였다. 또한, 학습 초기부터 강한 domain-adversarial gradient가 적용되어 학습이 불안정해지는 것을 방지하기 위해 GRL 강도를 학습 진행률에 따라 점진적으로 증가시키는 스케줄을 적용하였다. GRL 강도의 최댓값을 1.0으로 설정하여 도메인 분류 손실로부터 특징 추출기에 전달되는 그라디언트의 최대 크기가 원래 크기가 되도

표 5. Random seed에 따라 선택된 소스 도메인
Table 5. Source domains selected by random seed

Random Seed		Selected Categories
1	1	Cell_Phones_and_Accessories
	2	All_Beauty
	3	Video_Games
	4	Baby_Products
	5	Grocery_and_Gourmet_Food
	6	Health_and_Personal_Care
	7	Gift_Cards
	8	Sports_and_Outdoors
	9	Magazine_Subscriptions
	10	Amazon_Fashion
	11	Books
	12	Subscription_Boxes
	13	Pet_Supplies
	14	Appliances
	15	Office_Products
2	1	Baby_Products
	2	Beauty_and_Personal_Care
	3	Automotive
	4	Electronics
	5	Industrial_and_Scientific
	6	Toys_and_Games
	7	Office_Products
	8	Pet_Supplies
	9	Movies_and_TV
	10	Clothing_Shoes_and_Jewelry
	11	Cell_Phones_and_Accessories
	12	Home_and_Kitchen
	13	Books
	14	Health_and_Personal_Care
	15	Arts_Crafts_and_Sewing
3	1	Grocery_and_Gourmet_Food
	2	Cell_Phones_and_Accessories
	3	Electronics
	4	Subscription_Boxes
	5	Home_and_Kitchen
	6	Video_Games
	7	Magazine_Subscriptions
	8	Kindle_Store
	9	Automotive
	10	Sports_and_Outdoors
	11	Appliances
	12	Software
	13	Toys_and_Games
	14	Health_and_Household
	15	CDs_and_Vinyl

표 6. Random seed에 따라 선택된 타겟 도메인
Table 6. Target domains selected by random seed

Random Seed		Selected Categories
1	1	Arts_Crafts_and_Sewing
	2	Automotive
	3	Beauty_and_Personal_Care
	4	CDs_and_Vinyl
	5	Clothing_Shoes_and_Jewelry
	6	Digital_Music
	7	Electronics
	8	Handmade_Products
	9	Health_and_Household
	10	Home_and_Kitchen
	11	Industrial_and_Scientific
	12	Kindle_Store
	13	Movies_and_TV
	14	Musical_Instruments
	15	Patio_Lawn_and_Garden
	16	Software
	17	Tools_and_Home_Improvement
	18	Toys_and_Games
2	1	All_Beauty
	2	Amazon_Fashion
	3	Appliances
	4	CDs_and_Vinyl
	5	Digital_Music
	6	Gift_Cards
	7	Grocery_and_Gourmet_Food
	8	Handmade_Products
	9	Health_and_Household
	10	Kindle_Store
	11	Magazine_Subscriptions
	12	Musical_Instruments
	13	Patio_Lawn_and_Garden
	14	Software
	15	Sports_and_Outdoors
	16	Subscription_Boxes
	17	Tools_and_Home_Improvement
	18	Video_Games
3	1	All_Beauty
	2	Amazon_Fashion
	3	Arts_Crafts_and_Sewing
	4	Baby_Products
	5	Beauty_and_Personal_Care
	6	Books
	7	Clothing_Shoes_and_Jewelry
	8	Digital_Music
	9	Gift_Cards
	10	Handmade_Products
	11	Health_and_Personal_Care
	12	Industrial_and_Scientific
	13	Movies_and_TV
	14	Musical_Instruments
	15	Office_Products
	16	Patio_Lawn_and_Garden
	17	Pet_Supplies
	18	Tools_and_Home_Improvement

록 설정하였다.

DeepCORAL에서는 분류 손실과 CORAL 손실의 가중합을 사용하며, 이때 CORAL 손실의 반영 비율은 λ=0.1로 설정하였다.

MLDG의 경우, meta-train 도메인에서 가상 업데이트를 수행할 때 사용되는 inner learning rate를 1e-4로 설정하였다. 최종 목적 함수는 meta-train 손실과 meta-test 손실의 가중합으로 구성되며, 본 연구에서는 두 손실의 비중을 동일하게 반영하였다. meta-train 도메인에서 가상 업데이트를 수행한 후, 업데이트된 파라미터를 기반으로 meta-test 도메인의 손실을 계산하는 과정에서 2차 미분이 요구되지만 MLDG의 기반이 되는 MAML에서는 2차 미분항을 생략한 first-order approximation 방식이 완전한 2차 미분 기반 학습과 비교하여 성능 차이가 크지 않음을 보고하였다[28]. 이에, 본 연구에서도 계산 효율성을 고려하여 2차 미분을 생략한 first-order approximation 방식으로 MLDG를 구현하였다.

GroupDRO에서는 각 epoch가 시작할 때 도메인 가중치를 균등분포로 초기화하였다. 이후 epoch 내 미니 배치 단위로 계산된 도메인 손실을 기반으로 그룹 가중치가 갱신된다. 도메인 가중치 업데이트 식은 식 (13)과 같다.

$$q_g^{(t)} \leftarrow q_g^{(t-1)} \quad \exp(\eta L_g(\theta^{(t-1)})) \tag{13}$$

여기서 $q_g^{(t)}$는 t번째 업데이트에서 도메인 g에 할당된 가중치이고 $L_g(\theta^{(t-1)})$는 이전 업데이트 시점의 모델 파라미터 $\theta^{(t-1)}$에 대해 계산된 그룹 g의 손실이며, t는 배치 단위 업데이트 순서를 나타내는 인덱스이다. η는 도메인별 가중치 q의 업데이트 속도를 조절하는 하이퍼파라미터이고 본 연구에서는 η = 0.01로 설정하였다.

Mixup은 두 데이터가 혼합되는 비중을 λ로 조정할 수 있는데 이때 λ는 일반적으로 Beta 분포로부터 추출된다. 그러나, 텍스트 기반 Transformer 환경에서 Mixup 효과를 분석한 기존 연구[35]에서는 BERT 기반 모델에서 λ 값을 0.1부터 0.9까지 변화시키더라도 성능 차이가 크지 않음을 보고하였으며, 이에 따라 실험에서는 λ= 0.5를 고정하여 사용하였다. 본 연구에서도 이를 참고하여 λ 값을 0.5로 고정하여 실험을 수행하였다.

한편, MLDG와 GroupDRO는 다른 알고리즘과 달리 배치 구성 방식을 별도로 설정하였다. 일반 알고리즘에서는 학습 batch size를 32로 고정하였으나, MLDG와 GroupDRO에서는 한 배치에 포함될 도메인 수와 도메인별 샘플 수를 지정하고, 전체 배치 크기가 '도메인 수 × 도메인별 샘플 수'가 되도록 구성하였다. 다른 알고리즘과 동일하게 batch size를 32로 설정하고 15개의 소스 도메인을 사용할 경우, 한 배치에서 각 도메인에 포함되는 샘플 수는 약 2개 내외에 불과하다. 이

처럼 도메인별 샘플 수가 지나치게 적으면 각 도메인에서 충분한 샘플이 확보되지 않아, MLDG에서의 meta-train 도메인과 meta-test 도메인을 기반으로 한 메타 학습과 GroupDRO에서의 도메인별 그룹 손실 추정이 불안정해질 수 있다. 이러한 문제를 완화하기 위해 MLDG와 GroupDRO에 대해 배치에 포함될 도메인 수와 도메인별 샘플 수를 각각 8개와 4개로 설정하였다. 이를 통해 전체 배치 크기를 다른 알고리즘과 동일하게 유지하면서도 각 도메인에서 최소한의 샘플 수를 확보할 수 있도록 하였다. 다만 이러한 도메인 균형 배치 구성은 일반적인 랜덤 배치와 비교하여 도메인별 데이터 노출 빈도와 그라디언트 분산을 변화시킬 수 있으므로, MLDG와 GroupDRO의 성능 차이는 알고리즘의 목적함수뿐만 아니라 알고리즘 적용을 위해 사용한 배치 샘플링 방식의 영향도 일부 반영될 수 있다. 따라서, 본 실험은 목적함수 수준의 비교 실험이 아니라, 각 알고리즘의 학습 특성에 맞게 구성된 전체 학습 방식의 성능을 비교한 것이다.

V. 결 과

5-1 평가 및 결과

전체 33개 도메인 중 15개를 소스 도메인으로 사용하여 ERM과 5가지 도메인 일반화 알고리즘의 리뷰 유용성 예측 성능을 비교하였다. 표 7과 그림 1은 각 알고리즘의 미학습 타겟 도메인에 대한 평균 성능을 나타낸다. 실험 결과, Mixup이 가장 높은 평균 성능을 기록하였다. DeepCORAL과 ERM은 Mixup보다는 낮았지만, 유사한 수준의 성능을 보였으며, 두 방법 모두 전체 알고리즘 중 비교적 안정적인 결과를 나타냈다. 반면 MLDG, DANN, GroupDRO는 상위 세 가지 방법에 비해 상대적으로 낮은 성능을 기록하였으며, 특히 GroupDRO는 전체 알고리즘 중 가장 낮은 성능을 나타냈다.

알고리즘 유형별로 베이스라인인 ERM과 비교하면, 데이터 조작 기반 방법인 Mixup은 ERM보다 높은 성능을 보였고, 학습 전략 기반 방법인 MLDG와 GroupDRO는 ERM보다 낮은 성능을 보였다. 표현 학습 기반 방법인 DeepCORAL과 DANN은 서로 다른 성능 경향을 보였는데, DeepCORAL은 ERM보다 소폭 높은 성능을, DANN은 ERM보다 낮은 성능을 기록하였다.

기존 실험 결과가 특정 사전학습 언어모델에 의존하는지 확인하기 위해, 기존 BERT-base-uncased 외에 RoBERTa와 BART를 백본으로 사용한 추가 실험을 수행하였다. RoBERTa는 BERT와 유사한 인코더 기반 모델이지만 사전학습 방식이 다르며, BART는 인코더-디코더 구조를 사용하는 모델이므로, 서로 다른 사전학습 구조와 학습 방식에서도 결과가 유지되는지를 확인하기 위해 선정하였다. 모든 백본에서 동일한 데이터 분할과 알고리즘 구성을 적용하였으며, 각

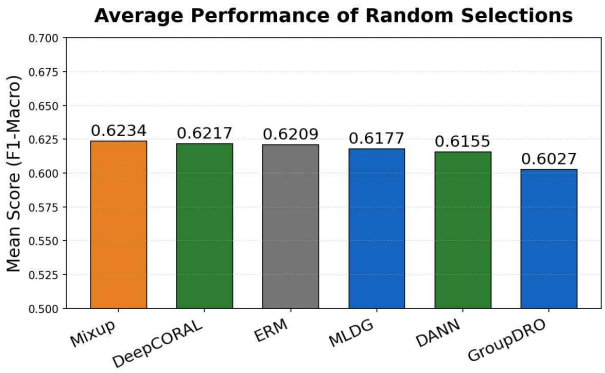


그림 1. 도메인 일반화 알고리즘별 평균 성능

Fig. 1. Average Performance by domain generalization algorithm

표 7. Random seed별 도메인 일반화 알고리즘의 평균 Macro F1-Score 성능

Table 7. Mean macro F1-score of domain generalization algorithms across random seeds

Algorithm	Seed 1	Seed 2	Seed 3	Avg±Std
Mixup	0.6261	0.6153	0.6287	0.6234±0.0071
Deep CORAL	0.6269	0.6085	0.6295	0.6217±0.0115
ERM	0.627	0.6066	0.629	0.6209±0.0124
MLDG	0.6187	0.6169	0.6174	0.6177±0.0009
DANN	0.6085	0.615	0.6232	0.6155±0.0074
Group DRO	0.6175	0.6049	0.5855	0.6027±0.0161

모델은 세 가지 도메인 분할 seed에서 평가하였다. 표 8에 나타난 것과 같이, 세 백본 모두에서 Mixup, DeepCORAL, ERM, MLDG, DANN, GroupDRO 순으로 동일한 평균 성능 순위가 나타났다. 이러한 결과는 기존의 알고리즘별 성능 양상이 BERT-base-uncased라는 특정 백본에만 한정되지 않고 RoBERTa와 BART에서도 대체로 일관되게 나타났음을

표 8. 백본 모델별 도메인 일반화 알고리즘 의 평균 Macro F1-score 성능

Table 8. Mean macro F1-scores of domain generalization algorithms across backbone models

Algorithm	BERT	RoBERTa	BART
Mixup	0.6234 ±0.0071	0.6241 ±0.0058	0.6204 ±0.0081
Deep CORAL	0.6217 ±0.0115	0.6208 ±0.0067	0.6200 ±0.0100
ERM	0.6209 ±0.0124	0.6194 ±0.0064	0.6195 ±0.0098
MLDG	0.6177 ±0.0009	0.6186 ±0.0036	0.6176 ±0.0043
DANN	0.6155 ±0.0074	0.6107 ±0.0084	0.6135 ±0.0051
Group DRO	0.6027 ±0.0161	0.6032 ±0.0075	0.6028 ±0.0276

보여주며, 백본 변화에 대한 결과의 강건성을 확인하는 근거로 해석할 수 있다.

5-2 결과 해석

앞선 결과에서 Mixup과 DeepCORAL은 ERM보다 평균 성능이 수치상 높게 나타난 반면, DANN, GroupDRO, MLDG는 ERM보다 낮은 평균 성능을 보였다. 그러나, 알고리즘 간 성능 차이는 전반적으로 크지 않았으므로, 평균 성능 차이가 통계적으로 유의한지 추가적으로 검정하였다.

본 연구에서는 random seed 분할에 따라 학습에 사용되는 소스 도메인 구성을 변경하였다. FSMT 평가 프로토콜에서는 미학습 타겟 도메인이 전체 도메인에서 소스 도메인을 제외한 집합으로 결정되므로, random seed에 따라 동일한 도메인이 미학습 타겟 도메인 또는 소스 도메인으로 사용될 수 있다. 이에 따라, 알고리즘별 도메인 수준 성능 차이를 산출하기 위해 먼저 해당 도메인이 미학습 타겟 도메인으로 포함된 각 분할에서 알고리즘과 ERM 간 Macro F1-score 차이를 계산하였다. 하나의 도메인이 여러 random seed에서 반복적으로 타겟 도메인으로 평가된 경우에는 해당 분할에서의 성능 차이를 평균하여 도메인 수준의 대표 성능 차이를 산출하였다. 예를 들어 특정 도메인이 random seed 1과 3에서는 미학습 타겟 도메인에 포함되고 random seed 2에서는 소스 도메인에 포함되었다면, random seed 1과 3의 성능만을 평균하여 해당 도메인의 대표 성능으로 사용하였다. 반대로 모든 random seed에서 소스 도메인으로 선택되어 미학습 타겟 도메인 성능을 계산할 수 없는 도메인은 통계 검정에서 제외하였다. 이렇게 산출한 도메인별 대표 성능 차이를 대상으로 Shapiro-Wilk 검정과 Wilcoxon signed-rank test를 수행하였다. 다만 도메인별로 미학습 타겟 도메인으로 평가된 횟수가 서로 다르며, 동일한 FSMT 분할에 포함된 여러 타겟 도메인의 성능은 동일한 학습 모델로부터 산출되므로 도메인 수준 성능값이 완전히 독립적이라고 보기에는 한계가 있다.

먼저 각 알고리즘과 ERM 간 도메인별 성능 차이값이 정규성 가정을 충족하는지 확인하기 위해 Shapiro-Wilk 정규성 검정을 수행하였다. 검정 결과, GroupDRO를 제외한 모든 알고리즘에서 p-value가 0.05 미만으로 나타나 정규성 가정을 충족하지 않았다. 이에 따라, 검정 절차의 일관성을 유지하기 위해 모든 알고리즘에 대해 비모수 검정인 Wilcoxon signed-rank test를 수행하였다. 또한, ERM과 다수의 알고리즘을 반복적으로 비교함에 따른 다중비교 문제를 고려하여 Wilcoxon signed-rank test의 p-value에 Holm 보정을 적용하였다.

Wilcoxon signed-rank test 결과, DeepCORAL과 Mixup은 ERM보다 평균 성능이 수치상 높게 나타났으나 그 차이는 통계적으로 유의하지 않았다. 따라서 DeepCORAL과 Mixup의 평균 성능 향상을 미학습 타겟 도메인 전반에 걸친 일관된 성능 개선으로 해석하기는 어렵다. 반면 DANN, GroupDRO, MLDG는 mean difference와 median difference가 모두 음수로 나타나 ERM보다 낮은 성능 경향을 보였다. 이는 리뷰 유용성 예측에서 단순한 도메인 일반화 기법의 적용만으로는 미학습 타겟 도메인에 대한 일반화 성능을 안정적으로 개선하기 어렵다는 점을 보여준다.

학습 전략 기반 접근 방식인 MLDG와 GroupDRO는 도메인 이동을 시뮬레이션하거나 worst-case group의 성능을 개선하도록 설계되었지만, 본 실험에서는 ERM 대비 성능 향상을 보이지 못하였다. 이는 리뷰 유용성 예측 환경에서 도메인 정보를 활용한 정교한 학습 전략이 반드시 일반화 성능 향상으로 이어지는 것은 아님을 시사한다.

반면, 데이터 조작 기반 방법인 Mixup은 전체 평균 성능에서 가장 높은 값을 보였다. Mixup은 도메인 정보를 직접 제거하기보다 학습 샘플과 라벨을 선형 보간하여 보다 부드러운 결정 경계를 학습하도록 한다. 이러한 특성이 특정 소스 도메인에 대한 과적합을 완화하는 데 기여한 것으로 해석할 수 있다. 그러나, Mixup의 평균 Macro F1-score는 0.6234로, DeepCORAL의 0.6217 및 ERM의 0.6209와 비교했을 때 그 차이는 매우 작았으며, ERM과의 차이도 통계적으로 유의하지 않았다. 따라서 이를 일반적인 성능 개선으로 해석하기보다는 제한적인 수치적 우위로 볼 필요가 있다.

표현 학습 기반 접근 방식 중 DeepCORAL은 ERM보다 높은 성능을 기록하였으나, 평균 차이는 0.0005에 불과했으

표 9. 도메인별 성능 차이(Algorithm - ERM)에 대한 Shapiro-Wilk 정규성 검정 결과

Table 9. Shapiro-Wilk normality test results for domain-wise performance differences (Algorithm - ERM)

Algorithm	Shapiro-Wilk statistic	p-value
DeepCORAL	0.6882	<0.005
DANN	0.9159	<0.005
GroupDRO	0.9538	0.1841
Mixup	0.6933	<0.005
MLDG	0.7830	<0.005

표 10. ERM 대비 알고리즘별 성능 차이에 대한 Wilcoxon signed-rank test 결과

Table 10. Wilcoxon signed-rank test results for performance differences relative to ERM

Algorithm	Mean Diff.	Median Diff.	W	p-value	Holm-adjusted p-value
Deep CORAL	0.0005	0.0001	263	0.9926	1.0000
DANN	−0.0074	−0.0068	129	0.0105	0.0315
Group DRO	−0.0200	−0.0203	54	< 0.001	<0.001
Mixup	0.0015	−0.0007	237	0.6247	1.0000
MLDG	−0.0039	−0.0055	124	0.0078	0.0312

며 통계적으로도 유의하지 않았다. 반면, 같은 표현 학습 기반 접근 방식인 DANN은 ERM보다 낮은 성능으로 나타났고 그 차이는 통계적으로 유의하였다. DANN은 적대적 학습을 통해 도메인 구분 정보를 제거하여 강한 도메인 불변 표현을 학습하는 방식이고 DeepCORAL은 도메인 간 공분산 차이를 최소화하여 특징 분포를 정렬하는 방식이다. 이러한 차이는 리뷰 유용성 예측에서 도메인 정보를 강하게 제거하는 방식보다, 도메인 간 분포 차이를 완화적으로 정렬하는 접근이 더 적합할 가능성을 시사한다. 특히, DANN의 낮은 성능은 여러 도메인으로의 일반화를 고려할 때도 도메인 특성이 단순히 제거해야 할 정보가 아닐 수 있음을 보여준다. 다만, DeepCORAL의 성능 개선 폭이 통계적으로 유의한 수준에 미치지 못하였으므로 리뷰 유용성 예측에서는 도메인 특화 정보를 완전히 제거하기보다, 필요한 정보를 유지하면서 도메인 간 분포 차이를 조정하는 방식이 더 적합할 수 있다는 해석을 일반화하기 위해서는 소스 도메인 수와 구성을 다양하게 설정한 추가 실험과 다른 리뷰 데이터셋을 활용한 검증이 필요하다.

Ⅵ. 논 의

6-1 시사점

본 연구는 다음과 같은 의의를 가진다.

리뷰 유용성 예측 문제를 도메인 일반화 관점으로 확장했다는 점에서 의의가 있다. 기존 리뷰 유용성 예측 연구는 대부분 특정 도메인 내에서 학습과 평가가 이루어지거나, 학습 데이터와 테스트 데이터가 유사한 분포를 따른다는 가정 아래 수행되었다. 반면, 본 연구는 상품 카테고리를 도메인으로 정의하고, 학습에 사용되지 않은 미학습 타겟 도메인에서의 성능을 평가하였다. 따라서, 본 연구는 리뷰 유용성 예측 문제를 단순한 분류 성능 향상 문제가 아니라, 새로운 상품 카테고리로의 일반화 문제로 확장했다는 점에서 의의를 가진다.

본 연구는 도메인 일반화 알고리즘의 적용이 항상 성능 향상으로 이어지는 것은 아님을 실증적으로 확인했다는 점에서 의의가 있다. Mixup과 DeepCORAL은 평균 성능에서 ERM보다 높은 값을 보였으나 그 차이는 통계적으로 유의하지 않았다. 또한, DANN, MLDG, GroupDRO는 ERM보다 낮은 성능 경향을 보였다. 이는 ERM이 도메인 일반화 알고리즘과 비교해도 충분히 경쟁력 있는 baseline으로 작동할 수 있음을 보여주며, 기존 연구[7],[8]의 결과와도 연결된다.

실험 결과, 동일한 표현 학습 기반 접근 방식 내에서 DeepCORAL은 ERM 대비 평균 성능이 향상된 반면 DANN은 오히려 하락하였다. 비록 DeepCORAL의 성능 향상은 통계적으로 유의하지 않았지만, 이러한 결과는 리뷰 유용성 예측에서 도메인 특화 정보를 무조건 제거하기보다 도메인 간 분포 차이를 조정하는 접근이 상대적으로 유리할 가능성을 시사한다. 따라서 도메인 일반화 알고리즘을 선택할 때는 도메인 정보를 처리하는 방식에 대한 고려가 필요하다는 점을, 실험을 통해 확인하였다.

6-2 한계점 및 향후 연구 방향

본 연구에는 몇 가지 한계가 있다.

첫째, 본 연구는 전체 33개 도메인 중 15개를 소스 도메인으로 사용하는 설정을 중심으로 실험을 수행하였다. 소스 도메인 수가 달라질 경우 알고리즘 간 상대적 성능과 일반화 양상이 달라질 가능성이 있다. 따라서, 도메인 수 K 를 다양하게 설정하여 소스 도메인 수에 따른 성능 민감도를 분석할 필요가 있다.

둘째, 본 연구에서는 Amazon 리뷰 데이터셋에서 상품 카테고리를 하나의 도메인으로 정의하고 실험을 수행하였다. 그러나, 실제 리뷰 환경에서 도메인 차이는 상품 카테고리뿐만 아니라, 플랫폼 또는 언어권 등 다양한 요인에 의해 발생할 수 있다. 따라서, 향후 연구에서는 상품 카테고리 외의 다양한 도메인 정의를 고려할 필요가 있다.

셋째, 본 연구에서는 유용성 투표 수가 1 이상인 리뷰만을 분석 대상으로 사용하였고, 도메인별로 균등한 개수로 샘플링하였다. 이러한 선별 방식은 유용성 투표가 없는 리뷰를 비유용 리뷰로 간주함으로써 발생할 수 있는 라벨 오류를 줄일 수 있으나, 신규 리뷰나 노출 기회가 적은 리뷰를 분석 대상에서 제외하므로 실제 플랫폼에 게시된 전체 리뷰의 분포를 완전히 반영하지 못할 수 있다. 따라서 본 연구의 모델은 최소 한 번 이상 유용성 평가를 받은 리뷰의 조건부 분포를 학습하며, 신규 리뷰나 노출이 적은 리뷰에 대한 일반화에는 한계가 있을 수 있다. 또한, 실제 Amazon 리뷰 데이터에서는 상품 카테고리별 리뷰 수에 상당한 차이가 존재하지만, 본 연구에서는 도메인 간 균형을 유지하기 위해 모든 카테고리에서 동일하게 3,000개의 리뷰를 샘플링하였다. 이로 인해 실제 플랫폼에서 나타나는 카테고리별 데이터 규모의 차이와 상대적 비중이 학습 데이터에 그대로 반영되지 않았다는 한계가 있다. 따라서 본 연구의 결과는 실제 리뷰 분포를 그대로 재현한 환경이라기보다, 도메인별 데이터 규모를 통제한 조건에서 알고리즘 간 일반화 성능을 비교한 결과로 해석할 필요가 있다. 향후 연구에서는 실제 카테고리별 데이터 규모와 분포를 유지한 조건에서도 결과가 일관되게 나타나는지 추가적으로 검증할 필요가 있다.

넷째, 도메인 분할에 따라 각 도메인이 미학습 타겟 도메인으로 평가된 횟수가 서로 달랐다. 한 번만 평가된 도메인의 성능은 특정 소스-타겟 분할의 영향을 상대적으로 크게 받을 수 있는 반면, 여러 분할에서 반복 평가된 도메인의 평균 성능은 분할에 따른 변동이 일부 완화될 수 있다. 본 연구에서는 각 도메인을 동일한 분석 단위로 사용하였으나, 이러한 평가 횟수의 차이가 도메인 수준 평균과 통계 검정의 안정성에

영향을 미쳤을 가능성이 있다. 향후 연구에서는 더 많은 도메인 분할을 사용하거나 각 도메인의 타겟 평가 횟수를 균등하게 구성할 필요가 있다.

다섯째, 본 연구에서는 각 도메인 일반화 알고리즘의 고유 하이퍼파라미터에 대한 체계적인 탐색을 수행하지 않고, 사전에 고정한 값을 모든 도메인 분할에 동일하게 적용하였다. 따라서 본 연구의 결과는 각 알고리즘을 개별적으로 최적화하여 얻은 최고 성능을 비교한 결과라기보다, 고정된 설정에서 대표적인 도메인 일반화 알고리즘을 리뷰 유용성 예측에 적용했을 때 나타나는 상대적인 성능 양상을 비교한 결과로 해석할 필요가 있다. 향후 연구에서는 미학습 타겟 도메인의 정보를 사용하지 않는 조건을 유지하면서 소스 도메인 검증 세트를 기반으로 동일한 탐색 예산과 일관된 선택 기준을 적용한 하이퍼파라미터 최적화를 수행할 필요가 있다.

Ⅶ. 결 론

본 연구는 리뷰 유용성 예측에서 상품 카테고리 간 도메인 이동 문제를 도메인 일반화 관점에서 다루고, 기본 학습 방식인 ERM과 대표적인 도메인 일반화 알고리즘을 동일한 조건에서 비교하였다. 이를 통해 도메인 일반화 알고리즘의 적용이 미학습 타겟 도메인 성능 향상으로 이어지는지와 알고리즘별 도메인 정보 활용 방식이 성능에 미치는 영향을 분석하였다.

실험 결과, Mixup이 가장 높은 평균 성능을 보였으며 DeepCORAL과 ERM도 유사한 수준의 성능을 기록하였다. 그러나, ERM과 각 알고리즘 간 성능 차이에 대한 통계 검정 결과, Mixup과 DeepCORAL의 평균 성능 향상은 통계적으로 유의하지 않았다. 반면 DANN, MLDG, GroupDRO는 ERM보다 낮은 성능 경향을 보였다. 이러한 결과는 본 연구의 설정과 리뷰 유용성 예측 태스크에서 도메인 일반화 알고리즘의 적용이 항상 ERM보다 우수한 성능을 보장하지는 않음을 보여준다. 또한, 도메인 일반화 성능은 알고리즘의 복잡도나 유형 자체보다 도메인 정보를 처리하는 방식에 영향받을 가능성이 있음을 시사하였다.

본 연구는 학술적으로 리뷰 유용성 예측 문제를 도메인 일반화 관점에서 분석함으로써, 기존의 동일 분포 기반 예측 연구를 미학습 타겟 도메인 일반화 문제로 확장했다는 의의를 가진다. 또한, 실제 온라인 플랫폼에서 새로운 상품 카테고리에 리뷰 유용성 예측 모델을 적용할 때, ERM을 우선 고려할 필요가 있음을 보여준다. 한편, Mixup은 ERM 대비 통계적으로 유의한 성능 향상을 보이지는 않았으나, 수치상 가장 높은 평균 성능을 기록하였다는 점에서 비교적 단순한 데이터 증강 방식의 활용 가능성을 탐색할 수 있다는 실무적 의의를 가진다. 향후 연구에서는 소스 도메인의 수와 구성, 도메인 선택 전략, 플랫폼 및 언어권과 같은 다양한 도메인 정의를 고려하여 리뷰 유용성 예측에서 도메인 일반화 성능에 영향을 미치는 요인을 보다 세밀하게 분석할 필요가 있다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2025-25425154, RS-2025-23525579).

참고문헌

[1] T. L. Ngo-Ye, A. P. Sinha, and A. Sen, “Predicting the Helpfulness of Online Reviews Using a Scripts-Enriched Text Regression Model,” *Expert Systems with Applications*, Vol. 71, pp. 98-110, 2017. https://doi.org/10.1016/j.eswa.2016.11.029

[2] Q.-C. He and Y.-J. Chen, “Dynamic Pricing of Electronic Products with Consumer Reviews,” *Omega*, Vol. 80, pp. 123-134, October 2018. https://doi.org/10.1016/j.omega.2017.08.014

[3] M. Bilal and A. A. Almazroi, “Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews,” *Electronic Commerce Research*, Vol. 23, No. 4, pp. 2737-2757, 2023. https://doi.org/10.1007/s10660-022-09560-w

[4] H. Lim, S. Park, Q. Li, X. Li, and J. Kim, “What Makes a Review Helpful? A Multimodal Prediction Model in ㄷe-Commerce,” *Electronic Commerce Research and Applications*, Vol. 76, 101586, 2026. https://doi.org/10.1016/j.elerap.2026.101586

[5] S. Saumya, P. K. Roy, and J. P. Singh, “Review Helpfulness Prediction on e-Commerce Websites: A Comprehensive Survey,” *Engineering Applications of Artificial Intelligence*, Vol. 126, 107075, 2023. https://doi.org/10.1016/j.engappai.2023.107075

[6] A. Y. K. Chua and S. Banerjee, “Analyzing Review Efficacy on Amazon.com: Does the Rich Grow Richer?,” *Computers in Human Behavior*, Vol. 75, pp. 501-509, October 2017. https://doi.org/10.1016/j.chb.2017.05.047

[7] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, ... and P. S. Yu, “Generalizing to Unseen Domains: A Survey on Domain Generalization,” *IEEE Transactions on Knowledge and Data Engineering,* Vol. 35, No. 8, pp. 8052-8072, August 2023. https://doi.org/10.1109/TKDE.2022.3178128

[8] I. Gulrajani and D. Lopez-Paz, “In Search of Lost Domain Generalization,” arXiv:2007.01434, July 2020.

https://doi.org/10.48550/arXiv.2007.01434

[9] A. J. Storkey and M. Sugiyama, “Mixture Regression for Covariate Shift,” in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 1337-1334, 2006.

[10] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, “Learning to Generalize: Meta-Learning for Domain Generalization,” in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans: LA, pp. 3490-3497, April 2018. https://doi.org/10.1609/aaai.v32i1.11596

[11] D. Yang, M. Zhang, Y. Ni, S. Xu, and K. Zhu, “A Winner Effect Based Approach for Multimodal Review Helpfulness Prediction with Deep Learning Algorithm,” *Knowledge-Based Systems*, Vol. 330, 114545, 2025. https://doi.org/10.1016/j.knosys.2025.114545

[12] Y. Xiao, H. Shao, S. Yan, J. Wang, Y. Peng, and B. Liu, “Domain Generalization for Rotating Machinery Fault Diagnosis: A Survey,” *Advanced Engineering Informatics,* Vol. 64, 103063, 2025. https://doi.org/10.1016/j.aei.2024.103063

[13] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, “A Brief Review of Domain Adaptation,” in *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, Las Vegas: NV, pp. 877-894, 2021. https://doi.org/10.1007/978-3-030-71704-9_65

[14] T. Zheng, Z. Lin, Y. Zhang, Q. Jiao, T. Su, H. Tan, ... and R. Law, “Revisiting Review Helpfulness Prediction: An Advanced Deep Learning Model with Multimodal Input from Yelp,” *International Journal of Hospitality Management*, Vol. 114, 103579, 2023. https://doi.org/10.1016/j.ijhm.2023.103579

[15] Y. Zhang and D. Zhang, “Automatically Predicting the Helpfulness of Online Reviews,” in *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, Redwood City: CA, pp. 662-668, August 2014. https://doi.org/10.1109/IRI.2014.7051953

[16] M. Olmedilla, M. R. Martínez-Torres, and S. Toral, “Prediction and Modelling Online Reviews Helpfulness Using 1D Convolutional Neural Networks,” *Expert Systems with Applications*, Vol. 198, 116787, 2022. https://doi.org/10.1016/j.eswa.2022.116787

[17] L. Kong, C. Li, J. Ge, V. Ng, and B. Luo, “Predicting Product Review Helpfulness—A Hybrid Method,” *IEEE Transactions on Services Computing*, Vol. 15, No. 4, pp. 2213-2225, July-August 2022. https://doi.org/10.1109/TSC.2020.3041095

[18] M. Fan, C. Feng, L. Guo, M. Sun, and P. Li, “Product-Aware Helpfulness Prediction of Online Reviews,” in *Proceedings of the World Wide Web Conference (WWW 2019)*, San Francisco: CA, pp. 2715-2721, May 2019. https://doi.org/10.1145/3308558.3313523

[19] X. Qu, X. Li, C. Farkas, and J. Rose, “An Attention Model of Customer Expectation to Improve Review Helpfulness Prediction,” in *Proceedings of the 42nd European Conference on Advances in Information Retrieval (ECIR)*, Lisbon, Portugal, pp. 836-851, 2020. https://doi.org/10.1007/978-3-030-45439-5_55

[20] X. Li, Q. Li, D. Ryu, and J. Kim, “A BERT-Based Review Helpfulness Prediction Model Utilizing Consistency of Ratings and Texts,” *Applied Intelligence*, Vol. 55, No. 7, 455, 2025. https://doi.org/10.1007/s10489-024-06100-x

[21] C. Chen, Y. Yang, J. Zhou, X. Li, and F. Bao, “Cross-Domain Review Helpfulness Prediction Based on Convolutional Neural Networks with Auxiliary Domain Discriminators,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans: LA, pp. 602-607, June 2018. https://doi.org/10.18653/v1/N18-2095

[22] C. Chen, M. Qiu, Y. Yang, J. Zhou, J. Huang, X. Li, and F. S. Bao, “Multi-Domain Gated CNN for Review Helpfulness Prediction,” in *Proceedings of the World Wide Web Conference (WWW 2019)*, San Francisco: CA, pp. 2630-2636, May 2019. https://doi.org/10.1145/3308558.3313587

[23] Y. Zhong, W. A. Zhou, and Z. Wang, “A Survey of Data Augmentation in Domain Generalization,” *Neural Processing Letters*, Vol. 57, No. 2, 34, 2025. https://doi.org/10.1007/s11063-025-11747-9

[24] F. Lu, Q. Tong, X. Jiang, X. Du, J. Xu, and J. Huo, “Prior Knowledge Embedding Convolutional Autoencoder: A Single-Source Domain Generalized Fault Diagnosis Framework Under Small Samples,” *Computers in Industry*, Vol. 164, 104169, 2025. https://doi.org/10.1016/j.compind.2024.104169

[25] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” arXiv:1710.09412, 2017. https://doi.org/10.48550/arXiv.1710.09412

[26] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, ... and V. Lempitsky, “Domain-Adversarial Training of Neural Networks,” *The Journal of Machine Learning Research*, Vol. 17, No. 59,

pp. 1-35, 2016.

[27] B. Sun and K. Saenko, “Deep CORAL: Correlation Alignment for Deep Domain Adaptation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, pp. 443-450, October 2016. https://doi.org/10.1007/978-3-319-49409-8_35

[28] C. Finn, P. Abbeel, and S. Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks,” in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, pp. 1126-1135, July 2017.

[29] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, “Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization,” arXiv:1911.08731, 2019. https://doi.org/10.48550/arXiv.1911.08731

[30] Y. Hou, J. Li, X. Fu, Z. He, A. Yan, X. Chen, and J. McAuley, “Bridging Language and Items for Retrieval and Recommendation: Benchmarking LLMs as Semantic Encoders,” in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, San Diego: CA, pp. 3251-3265, July 2026. https://doi.org/10.18653/v1/2026.acl-long.147

[31] R. Saptono and T. Mine, “Distribution-Adapted Model for Helpful Vote Prediction,” *IEEE Access,* Vol. 10, pp. 125194-125211, 2022. https://doi.org/10.1109/ACCESS.2022.3225558

[32] V. Vapnik, “Principles of Risk Minimization for Learning Theory,” in *Proceedings of the 5th International Conference on Neural Information Processing Systems,* Denver, Colorado, pp. 831-838, 1991.

[33] L. Cascone, L. Di Biasi, G. Genito, and M. Nappi, “DINO2-DR: A Trustworthy and Explainable Vision Transformer for Cross-Domain Diabetic Retinopathy Grading,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu: HI, pp. 595-604, 2025.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT),* Minneapolis: MN, pp. 4171-4186, June 2019. https://doi.org/10.18653/v1/N19-1423

[35] L. Sun, C. Xia, W. Yin, T. Liang, P. Yu, and L. He, “Mixup-Transformer: Dynamic Data Augmentation for NLP Tasks,” in *Proceedings of the 28th International Conference on Computational Linguistics (COLING),* Barcelona, Spain (Online), pp. 3436-3440, December 2020. https://doi.org/10.18653/v1/2020.coling-main.305



김유진(Yujin Gim)

2026년 : 한양대학교 에리카 광고홍보학과(문학사)
2026년 : 한양대학교 에리카 디자인테크놀로지전공 (공학사)
2026년～현 재: 한양대학교 휴먼컴퓨터인터랙션학과 석사과정
※관심분야 : 딥러닝(Deep Learning), 빅데이터(Big Data), 인간-컴퓨터 상호작용(Human-Computer Interaction, HCI) 등



박정은(Jeongeun Park)

2018년 : 연세대학교 정보대학원 (정보시스템학석사)
2024년 : 연세대학교 정보대학원 (정보시스템학박사)
2012년～2014년: KT 융합기술원
2014년～2017년: ㈜ 레어마켓
2024년～현 재: 한양대학교 휴먼컴퓨터인트랙션학과 조교수
※관심분야 : 딥러닝(Deep Learning), 사회기술 시스템(Socio-technical Systems), 상호작용 AI(Interactive AI), Explainable AI(XAI) 등