자연어처리 모델의 예측 편향, 어디에서 왔을까?

숭실대학교 ▌박건우*

1. 서 론

딥러닝 기술의 발달로 학습 기반의 예측 모델이 뛰어난 성능을 보여주고 있다. 하지만, 모델의 예측 과정에서 특정 그룹에 대한 차별 등 사회적으로 바람직하지 않은 패턴을 모델이 학습하고 재생산하는 문제가 발생하고 있다. 상용 얼굴 인식기의 성별 인식 성능이 피부색에 따라 달라지고[1], 재범 여부 판단 시 인종 정보가 중요한 영향을 미치며[2], 챗봇이 장애인 및 성 소수자 그룹에 대한 혐오 발언을 생산한다[3]. 인공지능 편향, 예측 편향, 불공정 예측 등 다양한 이름으로 불리는 이 문제는 분류[4], 언어 생성[5], 품사 태깅[6] 등 자연어처리(NLP)의 핵심 기술 및 응용에서 관측되고 있다.

언어는 인구통계적 정보, 성격 등 그 언어를 사용하는 사람의 속성을 반영한다[7]. 따라서, NLP 모델은 학습에 사용된 말뭉치에 담긴 관점을 직간접적으로 학습한다. 통계적 방법의 특성상 모델이 학습하는 편향을 완전히 없앨 수는 없을 것이며, 모델 예측이 문제를 발생시키지 않는다면 학습된 편향 자체로는 문제가 되지 않을 수 있다. 하지만, 모델의 예측으로 인해 특정 그룹에 대한 차별이 나타나는 등 편향된 예측을 만들게 된다면 문제가 된다[8]. 예를 들어, 한 기업의 인재채용 과정에 자기소개서 등의 텍스트 문서를 평가하는 자동 서류심사 모델을 도입한다고 하자. 만약 이 모델이 데이터 내 숨어있는 허위 상관관계를 학습하여 특정 성별, 연령대에 사용되는 언어 패턴을 이해하고 낮은 점수를 준다면 사회적, 법적 문제를 야기할 수 있을 것이다. NLP 모델의 책임감 있는 사용을 위해, 왜 모델 예측 편향이 발생하고 어떻게 예방할 수 있을지에 대한 체계적인 이해가 필요하다.

이 고에서는 예측 편향(Predictive bias)에 관한 Shah et al. (2020)의 정의를 소개하고[9], NLP 핵심 기술과 응용에 걸쳐 편향이 나타난 연구 사례를 다룬다. 더 나아가, 편향의 주요 원인 세 가지를 다루고, 각각에 대해 연구된 해결책들을 소개한다. 그림 1은 일반적인 지도학습 기반 NLP 파이프라인에서 나타날 수 있는 편향의 원인과 모델 예측으로 인해 나타나는 편향의 종류를 개괄하고 있다. 모델은 사전학습 임베딩, 학습 코퍼스 선택, 수작업 레이블링 등 NLP 작업의 여러 단계에서 편향을 학습할 수 있으며, 성별 등 데이터가 지니는 특정 속성값에 따라 불평등한 예측 결과를 만들 수 있다.

* 정회원

"이 연구는 정부의 재원으로 한국연구재단 및 정보통신기획평가원의 지원을 받아 수행되었음" (No. 2021R1F1A1062691, IITP-2022-RS-2022-00156360)

2. 예측 편향

NLP 모델에 대한 예측 편향은 다음과 같이 정의된다: 특정 응용에 대한 모델의 예측에 결과 차이(Outcome parity) 또는 오류 차이(Error disparity)가 나타난다면 예측 편향이 있다고 간주한다. 해당 정의는 일반적인 지도학습 기반 NLP 모델에 적용될 수 있으며, Shah et al. (2020)은 특정 응용에 따라 예측한 결과를 바탕으로 편향 여부를 판단해야 할 것을 강조하고 있다[9]. 즉, 동일한 NLP 모델은 어느 도메인에서 적용되는지에 따라 예측 편향을 가질 수도 가지지 않을 수도 있다.

2.1 결과 차이(Outcome disparity)

확률 변수 Y 는 모델 출력의 '참' 값, $\hat{Y}$ 는 모델 예측 값, A 는 데이터가 지니는 속성 또는 그룹을 나타낸다고 하자. 주어진 A에 대한 모델 예측 결과에 대한 분포 $Q(\hat{Y} \mid A)$ 와 속성 A에 대한 출력값의 이상적인 분포 $P(Y \mid A)$ 가 다를 때, 예측 모델에 '결과 차이'가 존재한다. 이상적인 분포 $P(Y|A)$ 는 목표하는 도메인 및 응용에 따라 달라질 수 있다. 모델 개발자는 도메인에 맞는 기준에 따라 이상적인 분포를 정의할 수 있다.

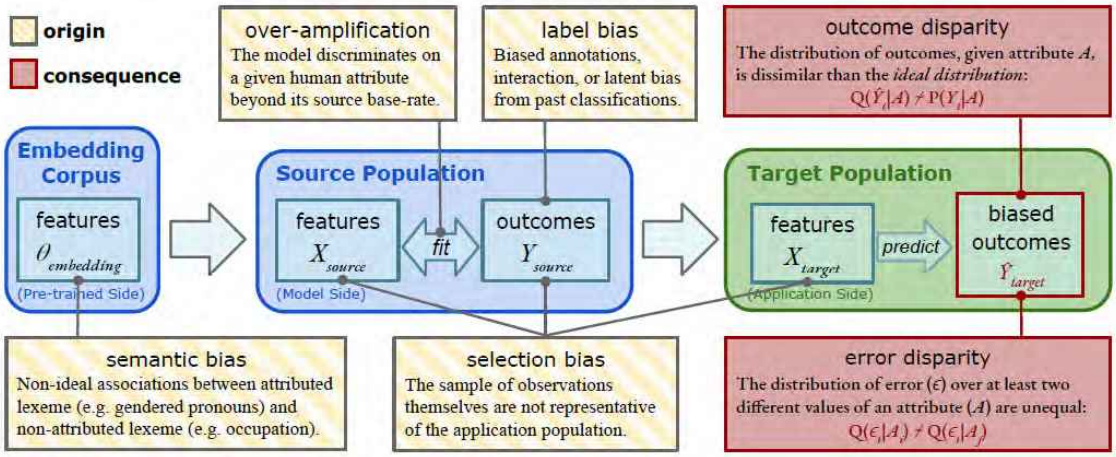

그림 1 NLP 모델의 예측 편향과 원인 (그림 출처: [9])

Zhao et al. (2017), Hendricks et al. (2018) 등은 이미지 캡셔닝[10], 의미 역할 라벨링[11] 등 작업에서 주어진 주어로 나타나는 성별에 따라 모델 출력의 분포가 달라질 수 있음을 보였다. 오븐에 대한 이미지를 입력으로 주었을 때는 여성, 스노보드에 대한 이미지는 남성을 주어로 생성하는 경향이 관측되었다. 입력 이미지의 종류에 따라 성별에 대한 분포가 동등할 것을 가정했을 때, 모델 예측이 이상적인 분포와 다른 '결과 차이'가 관측된 사례로 볼 수 있다.

2.2 오류 차이(Error disparity)

속성 A 값에 따라 모델 예측 오류의 패턴이 달라지는 경우 오류 차이가 있다고 한다. 오류는 모델 출력의 참값 Y 와 예측값 $\hat{Y}$ 차이의 크기로 정의된다. 속성 A 의 두 가지 다른 값에 대해 오류 차이가 관측된다면, 다시 말해, 특정 속성을 지닌 데이터에 대해 더 빈번한 오류가 관측된다면 주어진 모델은 오류 차이가 있다고 정의한다.

Sap et al. (2019)는 작성자의 인종에 따라 혐오 탐지 모델의 예측 결과가 달라질 수 있음을 보였다[4]. 흑인 사용자가 작성한 글에 대한 혐오 여부 판단 시 모델의 정확도가 낮게 나타났다. 송선영 등(2022)은 뉴스 인용구에 대한 감성 분석 사례 연구에서 기사를 발행한 언론사에 따라 감성 분석기의 정확도가 달라질 수 있음을 보였다[12]. 이는 특정 속성을 지닌 입력에 따라 성능이 달라지는 '오류 차이' 기반 예측 편향 사례이다.

3. 예측 편향의 원인과 해결책

어떤 NLP 모델이 편향된 예측 결과를 만들어 낸다면, 그 원인은 무엇이고 어떻게 해결할 수 있을까? 이 장에서는 세 가지 예측 편향의 원인과 해결책에 관한 최신 연구를 소개한다.

3.1 레이블 편향(Label bias)

많은 NLP 모델은 입력 텍스트와 정답값 간 매핑을 찾는 지도 학습 방법으로 학습된다. 특정 응용을 위해 지도 학습 NLP 모델을 구축하는 경우, 일반적으로 데이터에 대응되는 정답값을 얻기 위해 수작업 레이블링이 필요하다. 작업자(annotator)를 고용하여 작업자가 정답이라고 생각하는 레이블을 획득하고, 이를 학습을 위한 정답값으로 활용할 수 있다. 이때, 작업자의 판단이 잘못되었거나 특정 인구통계적 속성에 대한 사회적인 편견에 기반한다면, 예측 편향이 발생할 수 있다. 예를 들어, 아프리카계 미국인의 글을 모델이 혐오로 더 판단[4]하는 예측 편향의 원인은 작업자가 해당 인종에 대해 지닌 편견 때문일 수 있다. 더욱이, 과거에는 전문가를 고용하거나 소수의 훈련된 작업자를 레이블링 수작업에 이용하는 경우가 많았으나, 최근에는 저렴한 비용으로 수작업 레이블링을 할 수 있는 크라우드소싱[13] 방법이 널리 사용된다. Amazon Mechanical Turk 과 같은 크라우드소싱은 작업자 풀의 대표성이 제한되고 충분한 사전 훈련을 진행하기 어려워, 레이블 편향이 나타날 수 있다 [14].

작업자에 의해 발생하는 레이블 편향 중 가장 간단한 형태는 작업자의 부주의에 의해 발생한다. 이런 형태의 레이블 편향은 충분한 작업자 사전 훈련과 감독을 통해 예방할 수 있다. 더 파악하기 어려운 경우는 작업자가 지닌 사회적 고정관념과 같이 작업자의 잠재적인 속성이 레이블 품질에 영향을 주는 경우이다. 특히, 레이블링 작업이 작업자의 주관적 판단을 요구하는 경우[15]에 레이블 편향의 위험성이 더 커질 수

있다. Sachdeva et al. (2022)는 혐오 발언 레이블링에 작업자의 정체성이 영향을 미칠 수 있음을 문항반응이론(Item response theory)을 이용해 분석하였다[16].

NLP 시스템 구축을 위해 온라인 플랫폼에서 제공하는 정보를 레이블로 사용할 수도 있다. 예를 들어, Amazon Review Corpus[17]와 같이 사용자가 플랫폼에 남긴 리뷰 평점, 투표 점수 등을 지도학습을 위한 레이블로 사용할 수 있다. 이 방법은 앞서 언급한 수작업 레이블링의 비용 문제 등을 해결할 수 있다는 장점이 있으나, 사용자의 배경에 따라 플랫폼 내 정보에 대해 편향이 발생할 수 있음이 보고되었다. Park et al. (2018)은 고객 상담 라이브 챗 서비스 데이터의 만족도 설문 평가에 나타나는 레이블 편향을 분석하였다[18]. 레이블 데이터에서는 상담 만족도가 높은 사용자들이 많은 것으로 관측되었으나, 미 응답 데이터의 만족도를 추론하여 분석했을 때 불만족한 고객들이 더 많은 것으로 예측되었다. 즉, 서비스에 만족한 사용자들이 더 설문에 참여하여 레이블 데이터의 긍정 편향이 나타난 사례이다. Park et al. (2020)은 Amazon Review Corpus와 유사한 형태의 응답을 사용하는 Reddit의 투표 기록을 성향점수매칭(propensity score matching) 방법으로 분석[19]하여, 박사 학위를 지닌 사용자가 공유한 글은 일반 사용자가 업로드한 글에 비해 콘텐츠의 속성이 유사하더라도 높은 투표 점수를 받고 있음을 보였다. 만약, 높은 평가를 받는 글을 분류 예측하기 위해 텍스트 분류 모델을 학습한다면, 학위에 대한 사용자 속성과 레이블 간 허위 상관관계를 모델이 학습해 예측에 이용하는 편향이 발생할 수 있다.

작업자에 의한 레이블 편향은 작업자 간 일치도를 이용해 탐지할 수 있다[20]. Cohen's Kappa는 두 작업자 간 일치도를 측정하기 위해 사용할 수 있는 척도로, 두 작업자의 응답이 우연히 일치할 확률을 고려하여 관측된 일치도를 평가한다. 작업자가 3명 이상일 경우 Fleiss' Kappa를 사용할 수 있다. 어떤 수준의 일치도 값이 용인되는지는 도메인, 작업자 수, 가능한 클래스 수 등에 따라 달라질 수 있다. 두 명의 작업자가 두 개의 클래스에 대한 레이블링을 수행했을 때 기준으로, 0.2 이상이면 fair agreement, 0.4 이상이면 moderate agreement, 0.6 이상이면 substantial agreement를 보인다고 판단한다. 파일럿 태스크를 통해 품질이 높은 작업자를 선별하고, 작업 가이드라인을 상세히 기술하고, 충분한 사전 훈련을 통해 레이블 품질을 높일 수 있다. 하지만, 일치도가 높더라도 작업자 집단의 대표성이 낮을 때 사회적 고정관념 등으로 인한 레이블 편향이 발생할 수 있다.

3.2 선택 편향(selection bias)

선택 편향은 데이터의 대표성이 떨어지는 경우 발생한다. 구체적으로, 학습에 사용된, 즉 선택된 데이터의 분포가 모델이 실제로 적용될 데이터 분포와 다른 경우를 말한다. 선택 편향은 통제 실험, 설문조사 등 전통적인 연구방법을 적용할 때 연구 참여자를 모집하는 과정에서 발생할 수 있는 문제[21]로 사회과학 등 타 학문에서도 중요하게 다루어졌으며, 선택 편향이 있는 경우 피험자의 대표성이 떨어져 연구를 통해 얻은 결과의 신뢰성을 담보할 수 없게 된다. 유사하게, 자연어처리 모델이 대표성이 떨어지는 데이터를 학습하게 되면 모델이 예측해야 할 데이터 분포에서 나타나지 않은 입출력 관계에 의존해 편향된 예측 결과를 만들 수 있다.

선택 편향을 고려하여 NLP 모델을 만들 때 어려운 점은 어떤 인구통계적 정보를 통제하여야 하는지 불명확하다는 것이다. 일반적으로, 편향은 성별, 나이 등 그룹의 단일 속성에 의해 나타나지만, 여러 정체성이 결합되었을 때 새로운 형태의 편향이 나타날 수 있다. 이런 형태의 편향을 교차 편향(intersectional bias) 이라고 한다. Kim et al. (2021)은 혐오 발언 탐지 모델의 예측 패턴을 분석하여, 아프리카계 미국인 남성이 작성한 트윗의 경우 다른 집단의 작성자가 남긴 트윗에 비해 더 혐오 발언으로 판단되는 경향이 있음을 보였다[22]. 인종, 성별 등 하나의 요소만을 고려했을 때 해당 패턴이 보이지 않을 수 있어 교차 편향은 하나의 속성에 대해 나타나는 편향에 비해 더 다루기 어렵다.

데이터 선택 편향, 레이블 편향 등 데이터 셋이 내재하고 있는 예측 편향의 위험 요소를 완화시킬 수 있는 방법으로 다양한 문서화 전략이 제시되었다. 데이터셋 제작자가 데이터 구성을 위해 고려한 사항들을 상세히 기재하도록 하고, 데이터셋 사용자는 이를 바탕으로 NLP 모델의 예측 과정에서 나타날 수 있는 편향을 인지하도록 한다. Bender and Friedman (2018)은 NLP 기술 연구를 위해 Data Statement를 사용할 것을 제안하여[23], 데이터셋 구축의 목표, 수집 방법, 화자의 인구통계학적 속성 등을 상세히 기록하도록 하였다. 유사하게, Gebru et al. (2021)은 일반적인 머신러닝 데이터셋에 적용할 수 있는 문서화 프레임워크 Datasheets for Datasets을 소개하였고[24], 데이터셋 구축에 관한 동기, 데이터 구성, 수집 방법 등에

관한 57개 질문의 답을 기록하도록 하였다. 위와 같은 문서화 과정을 통해 사용자 뿐 아니라 제작자 또한 데이터셋에 존재하는 잠재적 편향을 보다 잘 이해할 수 있다는 점에서 그 의의가 있다. Mitchell et al. (2019)는 모델의 의도된 사용 맥락과 성능 측정에 관한 세부사항을 명시하도록 하여 모델 오용으로 인한 예측 편향을 완화시키는 문서화 방법 Model Cards를 제시[25]하였고, 간략한 버전의 모델 카드가 HuggingFace의 Model Hub에 도입되었다[26].

도메인 적응(domain adaptation) 기법을 적용하여 데이터의 선택 편향을 완화할 수 있다. Saunders and Byrne (2020)은 추론 단계에서 기계 번역 모델의 성별 편향을 감소시키기 위한 lattice scoring scheme을 제시하였다[27]. 송선영 등(2022)은 언론사 정보를 고려하여 리샘플된 데이터로 훈련된 감성 분석기가 분류 및 공정성 지표에서 모두 높은 성능을 보임을 보고하였다[12].

3.3 의미 편향(semantic bias)

단어 임베딩은 각 단어의 의미를 학습해 밀집 벡터로 표현하며, 유사한 의미를 지닌 단어가 벡터 공간 내 가까이 위치하도록 한다. 따라서, 벡터 연산을 통해 주어진 단어와 의미적으로 유사한 단어를 찾을 수 있을 뿐 아니라, 단어의 의미에 기반한 유추(analogy) 문제를 풀 수 있다. 예를 들어, “man is to king as woman is to x” 라는 문장에서 x에 들어갈 단어를 찾는 경우, 다음의 임베딩 벡터간 관계로부터 답이 queen 이 될 수 있다는 것을 알 수 있다: $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{king} - \overrightarrow{queen}$. 하지만, 임베딩 벡터간 관계를 통해 직업 등에 대한 성 고정관념이 드러날 수 있음이 보고되었다. 즉, “man is to programmer as woman is to x”에서 x에 들어갈 단어를 word2vec 임베딩 공간에서 살펴보면 x = “homemaker” 라는 결과를 얻을 수 있으며[28], 직업에 대한 성 고정관념을 임베딩 모델이 학습 및 표현한 사례이다. 이와 같이, 임베딩이 학습 데이터에 숨겨진 다양한 형태의 사회적 편향과 고정관념을 학습하고 표현하는 것을 의미 편향, 또는 임베딩 편향이라고 한다. 의미 편향은 word2vec, GloVe 등 정적(static) 임베딩 뿐 아니라 BERT 등 맥락을 고려한(contextualized) 임베딩에서도 관측되고 있다[29]. 최근 사용되는 대부분의 신경망 기반 NLP 모델에서 임베딩을 포함한다는 점을 고려했을 때, 의미 편향이 예측 편향의 주요 원인이 될 수 있다.

임베딩 표현에 내재된 편향을 측정하고 완화시키기 위한 많은 연구가 이루어졌다. Caliskan et al. (2017)은 사람이 지닌 암시적인 편향을 측정하기 위한 Implicit Association Test (IAT)에 기반하여, Word Embedding Association Test (WEAT)를 제안하였다[30]. IAT는 사람들이 지니고 있는 개념(예. ‘flowers’ or ‘insects’)과 속성(예. pleasantness and ‘unpleasantness’) 간 관계를 응답 시간을 이용해 측정한다. IAT에 기반해 아프리카계 미국인 이름을 부정적인 개념과 더 연관지어 생각하는 인지 편향이 발견되었다[31]. WEAT는 IAT와 유사하게, 두 개의 목표 단어 집합과 두 개의 속성 단어의 집합 간 임베딩 표현의 의미적 유사성 및 차이를 비교하여 의미 편향을 측정하며, IAT를 통해 관측되었던 인종 편향이 임베딩 공간에서 WEAT으로 재현됨을 보였다. 유사하게, WEAT를 문장 임베딩 레벨로 확장한 SEAT[32]가 있다.

WEAT와 SEAT는 정적 임베딩 공간에서 코사인 유사도를 이용해 편향을 측정하나, 맥락을 고려한 임베딩에서는 해당 방법이 일관적이지 않은 결과를 만들게 된다. 따라서, Kurita et al. (2019) 은 편향을 측정하고 싶은 속성을 포함하는 문장 템플릿을 만들고, masked LM 작업을 이용해 목표를 직접 예측하도록 하여 BERT 사전학습 언어모델이 지닌 편향을 측정하는 방법을 제안하였다[33]. 예를 들어, 프로그래머 직업에 대해 BERT 모델이 지니는 성별 편향을 측정한다고 하자. 이때, “[MASK] is a programmer”라는 템플릿을 이용해 관심 대상에 대한 확률 p_{target}을 추정하고, “[MASK] is a [MASK]”에 예측한 결과를 바탕으로 계산한 사전확률 p_{prior}로 정규화하여 관심 대상과 속성 간 연관성을 측정할 수 있다.

임베딩이 지닌 편향을 완화시키기 위한 대표적인 연구로 Bolukbasi et al. (2016)가 있으며[28], 성별 부분 공간을 찾아 편향을 없애거나 완화시키는 방법을 제안하였다. Ahn and Oh (2021)은 단일 언어 BERT에 나타난 인종 편향을 완화하기 위해 다국어 BERT를 사용하는 방법과 맥락 기반 alignment 기법을 제안하였다[34]. 텍스트 분류[35], 대화 생성[36], 기계 번역[37] 등 특정 다운스트림 작업에 초점을 맞춰 의미 편향을 감소시키기 위한 연구도 이루어졌다. 하지만, 편향이 완화된 임베딩을 다운스트림 작업에 적용하였을 때의 효과는 일관적이지 않은 것으로 알려져 있다.

임베딩 편향을 말뭉치에 존재하는 편향, 즉 특정 집단 혹은 사회의 편향을 측정하기 위한 하나의 도구로 사용할 수도 있다. Garg et al. (2018)은 Google Books Corpus 및 Corpus of Historical American English (COHA) 에 학습된 word2vec 단어 임베딩을 이용해 성별, 인종에 대한 편향이 지난 100년간 변화