

AI학습용 데이터의 편향성 검증 방안

AI시스템의 정확성과 공정성 확보를 위한
학습용 데이터의 편향성 검증, 어떻게 할 것인가?



정세린

- 데이터품질인증센터 인증심사팀 팀장
- 데이터품질인증기관 심사체계 조정 워킹그룹 멤버
- 초거대AI추진협의회 데이터 및 글로벌 확산 분과 위원
- 산업 인공지능 표준화 포럼 신뢰성 분과 위원

CONTENT

➤ 편향과 불공정

➤ AI학습용 데이터의 편향

➤ AI학습용 데이터의 편향 검증방법

➤ 향후 과제

소프트웨어 테스트 컨퍼런스

SOT&C 2024

AI · 데이터, 품질과 테스트 컨퍼런스

AI학습용 데이터의 편향성 검증 방안

 WISESTONE

편향과 불공정



편향

> 문자적 정의

한쪽으로 치우침

> 사회학 정의

특정 집단, 개인, 사건, 또는 생각에 대해 편견, 선입견, 고정관념 등에 기반하여 형성된 불균형적인 시각이나 태도를 의미

> 통계학 정의

추정량(Estimator)의 기댓값(Expected Value)과 실제 모수(Parameter) 값 사이의 차이를 의미

편향



> 인공지능 국제 표준에서의 정의

- 특정 객체, 사람 또는 그룹을 다른 것과 비교하여 체계적으로 다르게 처리하는 것 (ISO/IEC 22989 인공 지능 개념 및 용어)
- 대부분의 AI 시스템은 이러한 편향에 의존하여 분류, 군집화, 예측 또는 의사 결정을 함 (ISO/IEC 24027 시스템 및 AI 지원 의사 결정에서의 편향)
- 편향은 시스템의 목적에 따라 긍정적, 중립적, 부정적으로 평가 할 수 있음 (ISO/IEC 24027 시스템 및 AI 지원 의사 결정에서의 편향)

공정성



- 확립된 사실, 신념 및 규범을 존중하고, 편파나 부당한 차별에 의해 결정되지 않는 행동 또는 결과
- 문화, 세대, 지리 및 정치적 견해에 따라 달라짐
- 복잡하고, 매우 맥락적이며, 때로는 논쟁의 여지가 있는 개념
- 사회적, 윤리적으로 매우 맥락적인 특성을 가지므로 기준을 정의할 수 없음

출처 : ISO/IEC 24027 시스템 및 AI 지원 의사 결정에서의 편향

불공정



- **특정 그룹에 다른 그룹보다 우선적으로 이익을 주는 부당한 차별 대우**
- 불공정한 할당: AI 시스템이 기회를 부당하게 확대하거나 제한하여 일부 당사자에게 다른 당사자보다 부정적인 영향을 미치는 경우에 발생
- 불공정한 서비스 품질: AI 시스템이 일부 당사자에게는 다른 당사자보다 성능이 떨어지며, 기회나 자원이 확대되거나 제한되지 않더라도 발생
- 고정 관념화: AI 시스템이 기존 사회적 고정 관념을 강화하는 경우 발생
- 비하: AI 시스템이 경멸적이거나 모욕적인 방식으로 행동하는 경우에 발생

출처 : ISO/IEC 24027 시스템 및 AI 지원 의사 결정에서의 편향

편향과 불공정



- 편향은 공정성에 영향을 미칠 수 있는 많은 요소 중 하나임
- 편향된 입력이 항상 불공정한 예측과 행동을 초래하는 것은 아님
- 불공정한 예측과 행동이 항상 편향에 의해 발생하는 것은 아님

출처 : ISO/IEC 24027 시스템 및 AI 지원 의사 결정에서의 편향

의도된 편향



특정 목적을 위해 의도적으로
설계된 편향

대응 방안

- 투명성 제공
- 편향의 범위와 한계 명확히 안내
- 윤리적 검토 강화

의도되지 않은 편향



데이터나 알고리즘의 한계,
개발관계자들의 무의식적
편향으로 인해 발생하는 편향

대응 방안

- 데이터 편향성 검증 및 보완을 통해 편향 완화
- 편향 완화 알고리즘 적용
- 개발관계자들의 인지 편향 완화 방안 마련
- 지속적인 모니터링 및 평가

출처 : ISO/IEC 24027 시스템 및 AI 지원 의사 결정에서의 편향

의도되지 않은 편향

- AI 시스템의 공정성과 정확성을 저해하는 주요 요인
- 특정 집단에게 불리한 결과를 초래하여 사회적 불평등을 심화시키거나 개인에 피해를 줄 수 있음
- **AI학습용 데이터는 의도되지 않은 편향을 발생시키는 주요 요인**
- 시스템 개발 및 운영 과정에서 편향을 최소화하고 공정성을 확보하는 노력이 필수적

출처 : ISO/IEC 24027 시스템 및 AI 지원 의사 결정에서의 편향

소프트웨어 테스트 컨퍼런스

SOT&C 2024

AI·데이터, 품질과 테스트 컨퍼런스

AI학습용 데이터의 편향성 검증 방안

 WISESTONE

AI학습용 데이터의 편향



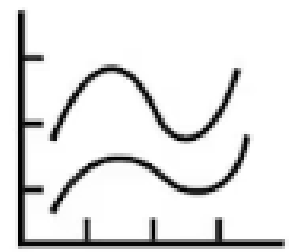
AI학습용 데이터 편향



- AI학습용 데이터가 데이터의 수집, 라벨링, 샘플링, 현실 반영 등으로 인해 한쪽으로 치우쳐 있는 것
- AI 학습용 데이터 편향은 AI 시스템의 공정성과 정확성을 위협하는 심각한 문제
- AI학습용 데이터 편향은 표본 편향과 표현 편향으로 구분할 수 있음

AI학습용 데이터 편향

표본 편향



데이터 셋이 모집단을 대표하지 못하는 경우 발생하는 편향

종류

대표성 편향(과대/과소), 선택 편향

표현 편향



데이터 자체에 내재된 편견이나 고정관념이 반영되는 경우 발생하는 편향

종류

측정 편향, 레이블 편향, 누락 데이터 편향, 차별적 언어 편향, 유해성 편향

AI학습용 데이터 편향 > 표본 편향

> 대표성 편향

특정 집단의 데이터가 과도하게 많거나 적어 해당 집단의 특성이 과장되거나 무시되는 편향

> 선택 편향

표본 선택 과정에서 특정 집단이 체계적으로 배제되거나 선호되는 편향

AI학습용 데이터 편향 > 표현 편향(1)

- > **측정 편향** 데이터 수집 또는 측정 과정에서 발생하는 오류로 인해 데이터가 왜곡되는 경우
- > **레이블 편향** 데이터 레이블링 과정에서 발생하는 오류 또는 주관적인 판단으로 인해 데이터가 잘못 분류되는 경우
- > **누락 데이터 편향** 특정 집단이나 특성을 가진 데이터가 누락되어 모델 학습에 사용되지 않는 경우

AI학습용 데이터 편향 > 표현 편향(2)

- > **차별적 언어 편향** 특정 집단에 대한 비하적 표현, 혐오 발언, 고정관념 등이 데이터에 포함되어 발생하는 편향

- > **유해 편향** 욕설, 폭력, 범죄, 자살 등 유해한 내용이 데이터에 포함되어 발생하는 편향

소프트웨어 테스트 컨퍼런스

SOTEC2024

AI·데이터, 품질과 테스트 컨퍼런스

AI학습용 데이터의 편향성 검증 방안

 WISESTONE



AI학습용 데이터의 편향 검증방법

표본 편향 검증 방법

- **데이터셋 분석** 인구통계학적 분포, 사회적 그룹 계층별 분포 시각화, 통계적 검정
- **교차 분석** 표본 선택 과정에서 특정 집단이 체계적으로 배제되거나 선호되는 경우를 확인하기 위해 집단 별로 교차하여 결과가 동일한지 분석

표현 편향 검증 방법(1)

- **이상치 탐지** 수치 데이터의 경우 이상치 데이터를 탐지
- **혼동 행렬 분석
(Confusion matrix)** 혼동 행렬의 수치를 분석하거나, 집단 간 혼동 행렬을 비교하여 수치의 동일성 확인
- **문장 및 단어 분석** 문장내 유해 단어나 표현이 있는지 확인하거나, 문장의 특정 집단에 대한 긍정/부정 비율 확인

표현 편향 검증 방법(2)

> 공정성 평가지표 적용

※ 공정성 평가지표 : 인공지능(AI) 시스템이 특정 집단이나 개인을 차별하지 않고 공정하게 작동하는지 평가하는 데 사용되는 척도이며, 공정성은 맥락에 따라 달라질 수 있으며, 특정 상황에 맞는 적절한 지표를 선택

인구통계학적 패리티 (Demographic Parity)	<ul style="list-style-type: none">○ 인구통계학적 범주 간 예측 비율이 동일한지 확인○ 예시: 대출 승인 모델에서 남성과 여성의 승인 비율이 전체 인구에서 남성과 여성의 비율과 동일해야 함
균등화 승률 (Equalized Odds)	<ul style="list-style-type: none">○ 인구통계학적 범주에 걸쳐 참긍정률(TPR)과 거짓긍정률(FPR)이 동일한지 확인
기회 균등 (Equal Opportunity)	<ul style="list-style-type: none">○ 인구통계학적 범주에 걸쳐 참긍정률(TPR)이 동일한지 확인○ 예시: 대출 상환 능력이 있는 사람들 중 저소득층과 고소득층이 대출 승인을 받을 확률이 동일해야 함
예측률 패리티 (Predictive Rate Parity)	<ul style="list-style-type: none">○ 인구통계학적 범주에 걸쳐 동일한 거짓긍정률(FPR)이 동일한지 확인○ 예시: 대출 승인을 받은 사람들 중 저소득층과 고소득층의 실제 대출 상환율이 동일해야 함

유용한 도구들 > 편향성

What-If Tool	Google에서 개발한 오픈 소스 도구로 모델의 편향을 시각적으로 분석하고 이해할 수 있도록 돕는 도구 https://pair-code.github.io/what-if-tool/
AI Fairness 360	IBM에서 개발한 오픈 소스 도구로, 데이터의 편향성을 감지하고 제거하는 기능을 제공 https://www.ibm.com/opensource/open/projects/ai-fairness-360/
TensorFlow Responsible AI Toolkit	TensorFlow에서 제공하는 툴킷 https://www.tensorflow.org/responsible_ai/fairness_indicators/tutorials/Fairness_Indicators_Example_Colab
SageMaker Clarify	아마존에서 개발한 서비스로, 코드를 작성하지 않고도 데이터 준비 중에 발생할 수 있는 편향을 식별 https://aws.amazon.com/ko/sagemaker/clarify/
Responsible AI	AI 개발자가 사람, 비즈니스 및 사회에 영향을 미칠 수 있는 위험과 피해를 식별하고 완화하는 데 도움이 되는 프레임워크로 공정성 https://azure.github.io/responsible-ai-hub/collections/

유용한 도구들 > 유해성

Google Jigsaw Toxicity Classifier	Google에서 개발한 오픈 소스 도구로, 텍스트 데이터의 유해성(예: 혐오 발언, 폭력적인 콘텐츠)을 감지하는 기능을 제공 https://jigsaw.google.com/the-current/toxicity/
Microsoft Azure Text Analytics	Microsoft에서 제공하는 클라우드 기반 서비스로, 텍스트 데이터의 유해성(예: 혐오 발언, 폭력적인 콘텐츠)을 감지하는 기능을 제공 https://learn.microsoft.com/en-us/azure/synapse-analytics/machine-learning/tutorial-text-analytics-use-mmlspark
Amazon Comprehend	Amazon에서 제공하는 클라우드 기반 서비스로, 텍스트 데이터의 유해성(예: 혐오 발언, 폭력적인 콘텐츠)을 감지하는 기능을 제공 https://aws.amazon.com/comprehend/
KSS(KISO Safeguard System) API Service	내 대표 포털 네이버와 카카오로부터 제공받은 욕설·비속어 DB를 활용해 개발한 API로, 해당 API는 약 80만 건의 욕설·비속어 DB를 활용해 입력된 표현 중 사전에 포함된 단어가 있는지 검사하고 그 결과를 필터링해 주는 서비스를 제공(유료) https://www.kiso.or.kr/%EC%9E%90%EC%9C%A8%EA%B7%9C%EC%A0%9C-db-%EC%95%88%EB%82%B4/kiso-safeguard-system/
MathLab VADER(Valence Aware Dictionary and sEntiment Reasoner) 알고리즘 활용	MathLab에서 제공하는 VADER 알고리즘을 사용하여 감성 분석 https://kr.mathworks.com/help/textanalytics/ug/analyze-sentiment-in-text.html

소프트웨어 테스트 컨퍼런스

SOT&C 2024

AI · 데이터, 품질과 테스트 컨퍼런스

AI학습용 데이터의 편향성 검증 방안

 **WISESTONE**



향후 과제 : 끊임없는 연구와 노력

비정형 데이터 편향(유해성) 측정 도구 연구 개발

- 문맥 기반 유해 표현 탐지 기술
- 유해 단어 목록 구축 및 사회적 합의
- 텍스트 뿐만 아니라, 이미지/ 동영상의 유해성 측정
도구 개발 필요



소프트웨어 테스트 컨퍼런스

SOT&C2024

AI·데이터, 품질과 테스트 컨퍼런스

AI학습용 데이터의 편향성 검증 방안

 **WISESTONE**

감사합니다

TRUST