Large Language Model에서의 인종 및 성별 편향 측정 연구

이주은[1], 배호[2,3]
[1]이화여자대학교 인공지능융합전공 석사과정
[2]이화여자대학교 사이버보안학과 교수
[3]이화여자대학교 인공지능융합전공 교수
jel@ewhain.net, hobae@ewha.ac.k

Research on Measuring Racial and Gender Bias in Large Language Model

Jueun Lee[1], Ho Bae[2,3]
[1]Dept. of Artificial Intelligence Convergence, Ewha Womans University
[2]Dept. of Cyber Security, Ewha Womans University
[3]Dept. of Artificial Intelligence Convergence, Ewha Womans University

요 약

Large Language Model(LLM) 사용이 증가하면서, LLM의 성별 및 인종에 대한 편향성은 사회적 불평등을 심화시킬 수 있는 중요한 문제로 대두되고 있다. 이에 LLM의 편향을 정확하고 신뢰성 있게 측정하는 도구가 필요하다. 본 논문은 LLM의 편향을 평가하는 방법론을 워드 임베딩 분석과 텍스트 생성 편향 분석으로 나누어 검토한다. 워드 임베딩 분석 방법은 단어 벡터 간 거리를 측정해 편향을 정량적으로 평가하는 방식으로, 간호사나 군인과 같은 단어들이 성별이나 인종과 같은 특정 집단과 얼마나 가깝게 매핑되는지를 분석하는 방식이다. 그러나 이 방법은 단어의 문맥적 의미 변화를 충분히 반영하지 못하는 한계가 있다. 반면, 텍스트 생성 편향 분석 방법은 LLM이 실제로 생성한 텍스트에서 나타나는 편향을 직접 평가하는 방식이다. 이를 위해 연구자는 성별이나 인종과 관련된 편향이 드러날 수 있는 문장들로 데이터셋을 구성하고, LLM이 이를 어떻게 처리하는지 분석한다. 이 방법은 문맥을 반영해 모델이 생성한 텍스트에서 편향을 평가할 수 있다는 장점이 있지만, 연구자가 데이터셋을 구축하는 과정에서 주관적 판단이나 편향이 개입될 가능성이 있으며, 평가할 수 있는 시나리오가 제한적이라는 한계가 있다. 본 논문은 이러한 한계를 극복하기 위한 향후 연구로, 합성 데이터를 활용하여 데이터셋을 구축하고, 이를 통해 텍스트 생성 편향을 분석하는 방법을 제안한다. 합성 데이터는 다양한 시나리오를 기반으로 무한히 생성할 수 있어, 특정 시나리오에 제한되지 않고 LLM의 편향을 폭넓게 평가할 수 있다. 또한 연구자의 개입을 줄여 데이터셋 구축 시 발생할 수 있는 편향을 최소화하고, 더 공정하고 신뢰성 있는 평가를 가능하게 한다. 이에 따라 합성 데이터를 이용한 텍스트 생성 편향 분석 방법은 LLM의 성별 및 인종 편향을 보다 객관적으로 평가하는 도구로서 중요한 역할을 할 것으로 기대한다.

1. 서론

Large Language Model (LLM)[1]은 방대한 양의 데이터를 학습하여 사람과 유사한 수준의 텍스트 생성 및 이해를 할 수 있는 능력을 갖추고 있는 생성형 인공지능이다. LLM은 복잡한 문제 해결 능력, 인간 언어 모방 능력 등으로 금융, 의료, 법률 등의 산업에서도 적극 연구 및 활용되고 있다[2].

그러나 이러한 기술적 이점에도 불구하고, LLM이 과거의 편향된 데이터를 학습함으로써 성별과 인종에 대한 편향된 결과를 생성할 위험이 존재한다[3].

인종 편향은 특정 인종에 대한 부정적이거나 고정관념에 기반한 결과를 생성하는 현상을 의미한다. 이는 LLM이 훈련된 데이터에 내재된 인종적 불균형이나 차별적 표현을 학습했기 때문에 발생한다[4].

성별 편향은 남성 혹은 여성에 대한 고정된 역할이나 특성을 강조하는 잘못된 결과를 내포하는 것으로, 성 역할에 대한 사회적 편견이 반영된 데이터를 학습한 결과이다[5].

이러한 편향이 그대로 방치될 경우, 이는 다양한 사회적 문제를 초래할 수 있다. 예를 들어, 인종 편향은 특정 인종에 대한 부정적 인식을 강화하고 차별을 조장할 수 있으며, 성별 편향은 성차별적인 고정관념을 통해 성 평등에 역행할 수 있다[6]. 따라서 LLM에서 발생하는 편향을 해결하기 위한 연구는 중요하며, 마찬가지로 LLM의 편향을 측정하는 기술적 도구의 정확성과 신뢰성, 공정성을 확보하는 것

은 필수적인 과제로 자리 잡고 있다[7].

이에 LLM의 편향을 측정하는 다양한 연구들이 활발하게 진행되고 있다. 이러한 LLM 편향을 측정하는 연구는 크게 워드 임베딩 분석과 텍스트 생성 편향 분석으로 나뉜다.

본 논문에서는 LLM의 인종 및 성별 편향을 중심으로 편향을 측정하는 도구에 대한 최신 연구 동향을 워드 임베딩 분석 방식, 텍스트 생성 편향 분석 방식으로 나누어 조사한다. 또한, 이를 바탕으로 LLM 편향 측정 분야에서 진행되어야 할 앞으로의 연구 방향을 제시하고자 한다.

2. 본론

2.1. 워드 임베딩 분석을 통한 LLM 편향 측정

워드 임베딩 분석은 단어 간 벡터 거리를 측정해 특정 단어들이 인종, 성별 등의 고정관념에 따라 임베딩 거리상 얼마나 가깝게 매핑되는지를 분석하는 방법이다. 이러한 방식으로 편향을 측정한 연구로는 [8], [9], [10], [11], [12], [13], [14], [15] 등이 있다.

대표적으로 2017년, Caliskan, A. et al.[9]은 자연어 처리 모델에 인간 사회의 편향이 학습된다는 점을 처음으로 실증하였다. 구체적으로, 성별, 인종에 따른 고정관념이 모델에 어떻게 반영되는지를 단어 간 임베딩 거리를 통하여 측정하는 방식인 Word Embedding Association Test (WEAT)를 제안하였다. 이를 통해 '남성'이라는 단어는 '과학자'와 같은 직업과 더 가까이 연결되는 반면, '여성'이라는 단어는 '간호사' 등과 더욱 가까이 연결되어 있다는 사실을 증명하며 편향을 입증하였다. 그러나 이는 정적 임베딩으로만 측정한다는 점에서 문맥에 따른 평가가 반영되지 않았다는 점, 이에 평가하는 시나리오의 다양성이 부족하다는 점 등의 한계가 있다.

해당 논문의 한계를 극복하고자 수많은 연구가 추가적으로 이루어졌으며, 그중 대표적인 예로 2019년 발표된 May, Chandler, et al.[10]이 있다. 이는 WEAT가 단어 수준에서만 편향을 측정한다는 점을 극복하고자 문장 수준 편향을 측정하는 Sentence Encoder Association Test (SEAT)를 제안하였다. 그러나 SEAT는 7개의 문장 인코더를 사용할 뿐, 실질적으로는 정적 단어 임베딩 방식과 유사한 평가 방식을 따르고 있어 문맥에 따른 세밀한 편향을 충분히 반영하지 못할 수 있다는 지적을 받았다.

이를 개선하기 위한 연구들이 진행되면서 2024년엔 Dobrzeniecka et al.[13]이 베이지안 접근법을 도입하여 단어 임베딩의 편향 측정에 내재된 불확실성을 고려한 방법을 제안하였다. 기존 WEAT나 SEAT와 같은 단일 지표 기반의 방법들과 달리 Google, GloVe, Reddit과 같은 다양한 임베딩 모델에서 베이지안 분석을 적용하여 편향을 측정하였다. 기존 정적 임베딩의 편향 측정 방식에 대한 보완책을 제시하였다는 점에서 의의를 갖지만, 실험의 범위와 시나리오의 다양성 부분에서는 여전히 한계가 남아 있다. 이에 실생활 시나리오와 같은 복잡한 환경에서의 추가적인 검증과 실험이 필요하다.

2.2. 텍스트 생성 편향 분석

텍스트 생성 편향 분석은 모델이 실제로 생성한 텍스트에서 성별, 인종 등과 관련된 편향이 어떻게 나타나는지를 평가하는 방법이다. 이 방식은 실생활 시나리오 및 실질적 결과를 분석하기 어렵다는 워드 임베딩 분석의 한계를 보완할 수 있다. 주요 연구로는 [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27] 등이 있다.

대표적으로 2018년, Zhao, Jieyu, et al.[16]은 LLM이 실제로 생성한 문장에서의 성별 편향을 평가할 수 있는 3,160개의 문장 데이터셋, Winobias를 제안하였다. Winobias는 성별 고정관념을 반영한 문장쌍을 통해, 특정 직업군을 지칭하는 대명사가 어느 성별을 가리키는지를 평가하여 편향을 측정한다. 그러나 이는 46개의 직업군만을 대상으로 3,160개의 시나리오를 평가하기에, 실제 편향이 발생할 수 있는 시나리오를 모두 반영하기엔 매우 제한된 범위에서 적은 수의 시나리오만을 분석하고 있다는 한계를 갖고 있다. 또한, 성별에만 특화되어 편향을 분석하고 있다는 한계가 있다.

이러한 한계를 극복하고자 해당 논문을 기점으로 다양한 시나리오를 포함한 데이터셋을 설계함으로써 LLM의 편향을 평가하고자 하는 연구가 활발히 진행되었다[17], [18], [19], [20], [21], [26]. 대표적으로 2021년 Dhamala, Jwala, et al.[21]은 Bias in Open-Ended Language Generation Dataset (BOLD)라는 성별, 인종 등의 다양한 도메인에서 편향을 측정할 수 있는 23,679개의 문장 데이터셋을 구축하여 제안하였다. 그러나 BOLD는 Wikipedia 문서에서 특정 규칙에 따라 추출된 문장을 기반으로 데이터셋을 구축하여, 실생활에서의 사회적 상황이나 복잡한 맥락을 반영하지 못하였다는 점, 사용하는 프롬프트 대부분이 6단어에서 9단어로 구성된 짧

은 문장들이기 때문에 모델이 실제로 더 긴 대화나 복잡한 맥락에서 어떻게 편향을 나타내는지 평가하기 어렵다는 점 등의 한계가 있다.

이후 이러한 면을 개선하기 위한 연구들이 진행되면서[22], [23], [24], [25], [27], 2024년 Wan, Yixin, and Kai-Wei Chang.[27]은 실생활에서 성별, 인종이 복합적인 편향을 불러일으킬 수 있음에도 성별과 인종을 별개의 편향으로 평가하던 이전 연구들을 지적하며 성별, 인종 각각의 편향은 물론이고 성별과 인종의 교차점에서 발생하는 편향을 분석하였다. 이에 LLM이 백인 남성은 리더로, 흑인 여성은 도우미로 묘사하는 경향이 있음을 밝혀냈다. 그러나 기존 데이터셋을 기반으로 교차적 편향을 분석한 것이기에 새로운 데이터셋 구축은 이루어지지 않았다. 따라서 이 연구는 실생활 시나리오에서 교차적 편향을 확장해 분석한 데 의의가 있으나, 실생활의 다양한 시나리오를 충분히 포괄하지는 못하였다.

2.3. 향후 연구 방향

LLM에서 성별, 인종 편향과 같은 편향을 분석하는 연구는 활발히 이루어져 왔다. 그러나 여전히 해결해야 할 과제가 남아있다.

현재 해결해야 할 가장 주요한 과제는, LLM 편향을 측정하기 위한 도구에 사용될 데이터셋이 제한된 시나리오만을 포함하고 있다는 점이다. 현재는 특정 도메인이나 제한된 범위의 시나리오에서 LLM을 평가하고 있다. 그러나 편향은 복잡하고, 다양한 모습으로 나타나며, 아직 정의되지 않은 모양으로도 나타난다. 그러나 현재 연구들은 주로 Wikipedia, 뉴스 등의 출처에서 데이터를 수집한 후, 이를 인간이 직접 정제하거나 기존에 발견한 시나리오에 의존하여 일부 응용하는 방식으로 데이터셋을 구축해왔다. 이로 인해 데이터셋이 포함하는 시나리오의 수는 제한적일 수밖에 없으며, 해당 데이터셋을 구축한 연구자의 편향이나 경험이 반영될 가능성도 존재한다. 결국, 편향 분석은 이 제한된 시나리오 안에서만 이루어질 수밖에 없다는 한계를 가진다.

따라서 합성 데이터를 활용한 대규모 데이터셋 구축 연구가 필요하다. 합성 데이터는 다양한 시나리오를 이론상 무한히 생성할 수 있어, 기존의 제한된 시나리오에서 벗어나 더 폭넓고 다양한 편향을 분석할 수 있게 한다. 이를 통해 LLM의 편향을 더 깊이 연구하고, 복잡한 상황을 반영한 데이터셋으로 편향 분석의 정확성과 포괄성을 높일 수 있을 것으로 기대한다.

3. 결론

LLM에서 발생하는 성별 및 인종 편향을 분석하는 연구는 크게 워드 임베딩 분석과 텍스트 생성 편향 분석이 있다.

워드 임베딩 분석은 수학적 증명이 가능해 더욱 정량적으로 평가할 수 있다는 이점이 있으나, 실제 LLM이 생성한 텍스트 내 분석이 어려워, 최근에는 텍스트 생성 편향 분석이 주를 이루고 있다.

이처럼 텍스트 생성 편향 분석을 위한 데이터셋 구축과 평가 방법론이 활발히 연구되고 있지만, 데이터셋의 범위가 제한적이고 연구자의 편향이 반영될 가능성이 있다.

이러한 한계를 극복하기 위해 앞으로는 더욱 다양하고 객관적인 시나리오에서의 LLM 편향에 대한 평가가 가능하도록 합성 데이터를 활용한 대규모 데이터셋 구축과 이에 대한 연구가 필요하다.

4. 사사

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-02068, 인공지능 혁신 허브 연구개발)

참고문헌

[1] Brown, Tom B. "Language models are few-shot learners."arXiv preprint arXiv:2005.14165(2020).

[2] Bommasani, Rishi, et al. "On the opportunities and risks of foundation models."arXiv preprint arXiv:2108.07258(2021).

[3] Bender, Emily M., et al. "On the dangers of stochastic parrots: Can language models be too big?□."Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021.

[4] Blodgett, Su Lin, et al. "Language (technology) is power: A critical survey of" bias" in nlp."arXiv preprint arXiv:2005.14050(2020).

[5] Zhao, Jieyu, et al. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints."arXiv preprint arXiv:1707.09457(2017).

[6] Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning."ACM computing surveys (CSUR)54.6 (2021): 1-35.

[7] Sun, Tony, et al. "Mitigating gender bias in natural language processing: Literature review."arXiv preprint arXiv:1906.08976(2019).

[8] Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases."Science356.6334 (2017): 183-186.

[9]Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings."Advances in neural information processing systems29 (2016).

[10] May, Chandler, et al. "On measuring social biases in sentence encoders."arXiv preprint arXiv:1903.10561(2019).

[11] Ethayarajh, Kawin, David Duvenaud, and Graeme Hirst. "Understanding undesirable word embedding associations."arXiv preprint arXiv:1908.06361(2019).

[12] Rakivnenko, Vasyl, et al. "Bias in Text Embedding Models."arXiv preprint arXiv:2406.12138(2024).

[13] Dobrzeniecka, Alicja, and Rafal Urbaniak. "A Bayesian approach to uncertainty in word embedding bias estimation."Computational Linguistics(2024): 1-55.

[14] Rai, Rohit Raj, and Amit Awekar. "Effect of dimensionality change on the bias of word embeddings."Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD). 2024.

[15] Freestone, Matthew, and Shubhra Kanti Karmaker Santu. "Word Embeddings Revisited: Do LLMs Offer Something New?."arXiv preprint arXiv:2402.11094(2024).

[16] Zhao, Jieyu, et al. "Gender bias in coreference resolution: Evaluation and debiasing methods."arXiv preprint arXiv:1804.06876(2018).

[17] Kiritchenko, Svetlana, and Saif M. Mohammad. "Examining gender and race bias in two hundred sentiment analysis systems."arXiv preprint arXiv:1805.04508(2018).

[18] Nadeem, Moin, Anna Bethke, and Siva Reddy. "StereoSet: Measuring stereotypical bias in pretrained language models."arXiv preprint arXiv:2004.09456(2020).

[19] Nangia, Nikita, et al. "CrowS-pairs: A challenge dataset for measuring social biases in masked language models."arXiv preprint arXiv:2010.00133(2020).

[20] Gehman, Samuel, et al. "Realtoxicityprompts: Evaluating neural toxic degeneration in language models."arXiv preprint arXiv:2009.11462(2020).

[21] Dhamala, Jwala, et al. "Bold: Dataset and metrics for measuring biases in open-ended language generation."Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021.

[22] Wan, Yixin, et al. ""kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters."*arXiv preprint arXiv:2310.09219*(2023).

[23] Kotek, Hadas, Rikker Dockum, and David Sun. "Gender bias and stereotypes in large language models."*Proceedings of the ACM collective intelligence conference*. 2023.

[24] Zhao, Jinman, et al. "Gender Bias in Large Language Models across Multiple Languages."*arXiv preprint arXiv:2403.00277*(2024).

[25] Rhue, Lauren, Sofie Goethals, and Arun Sundararajan. "Evaluating LLMs for Gender Disparities in Notable Persons."*arXiv preprint arXiv:2403.09148*(2024).

[26] Wang, Ze, et al. "JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models."arXiv preprint arXiv:2406.15484(2024).

[27] Wan, Yixin, and Kai-Wei Chang. "White Men Lead, Black Women Help: Uncovering Gender, Racial, and Intersectional Bias in Language Agency."arXiv preprint arXiv:2404.10508(2024).