제로샷 분류를 활용한 성별 편향 완화 성별 예측 방법

김연희 [1], 최병주 [2], 김종길 [3]
[1] 이화여자대학교 인공지능융합전공 박사과정
[2] 이화여자대학교 컴퓨터공학과 교수
[3] 이화여자대학교 사이버보안학과 교수

heedong@ewha.ac.kr, bjchoi@ewha.ac.kr, jongkil@ewha.ac.kr

Gender Bias Mitigation in Gender Prediction Using Zero-shot Classification

Yeonhee Kim[1], Byoungju Choi[2], Jongkil Kim[3]
[1]Dept. of Artificial Intelligence Convergence, Ewha Womans University
[2]Dept. of Computer Science and Engineering, Ewha Womans University
[3]Dept. of Cyber Security, Ewha Womans University

요 약

자연어 처리 기술은 인간 언어의 이해와 처리에서 큰 진전을 이루었으나, 학습 데이터에 내재한 성별 편향이 모델의 예측 정확도와 신뢰성을 저하하는 주요한 문제로 남아 있다. 특히 성별 예측에서 이러한 편향은 더욱 두드러진다. 제로샷 분류 기법은 기존에 학습되지 않은 새로운 클래스를 효과적으로 예측할 수 있는 기술로, 학습 데이터의 제한적인 의존성을 극복하고 다양한 언어 및 데이터 제한 상황에서도 효율적으로 작동한다. 본 논문은 성별 클래스 확장과 데이터 구조 개선을 통해 성별 편향을 최소화한 새로운 데이터셋을 구축하고, 이를 제로샷 분류 기법을 통해 학습시켜 성별 편향성이 완화된 새로운 성별 예측 모델을 제안한다. 이 연구는 다양한 언어로 구성된 자연어 데이터를 추가 학습하여 성별 예측에 최적화된 모델을 개발하고, 제한된 데이터 환경에서도 모델의 유연성과 범용성을 입증한다.

1. 서론

자연어 처리(Natural Language Processing, NLP) 기술은 인간 언어의 컴퓨터 이해와 처리를 목표로 하며, 기계학습 및 딥러닝의 발전을 통해 큰 진전을 이루었다.[1] 이러한 기술의 발전에도 불구하고, 대부분의 NLP 연구는 영어 중심의 데이터에 치중되어 있으며, 이는 다양한 언어에 대한 연구의 부족으로 이어져 비영어권 언어의 연구 개발에 큰 공백을 남긴다[2][3]. 또한, 학습 데이터에 내재한 편향은 모델의 성능에 부정적인 영향을 미치며, 이는 특히 성별 예측에서 문제가 두드러진다[4][5]. 현재의 성별 예측 모델은 대부분 '여성'과 '남성'의 이분법적 분류에 의존하고 있어, 성별을 예측할 수 없는 경우를 반영하지 못하며, 이에 따라 성별 예측의 정확도와 신뢰성이 저하된다. 본 연구는 제로샷 분류 기법을 활용하여 성별 편향을 최소화하는 동시에 다양한 언어 환경에서의 모델 적용성을 탐구함으로써, 성별 예측 모델의 정확도와 포괄성을 향상하는 새로운 접근 방법을 제안한다. 이를 통해 제한된 데이터셋에서도 효과적으로 작동할 수 있는 성별 예측 모델을 개발하고, 모델의 범용성을 검증하는 것을 목표로 한다.

2. 관련연구

2.1. NLP 에서 성별 예측에 관한 연구

성별 예측은 NLP 분야에서 중요한 연구 중 하나로, 이에 따라 다양한 기법이 연구되고 있다. 성별 예측은 주로 개인의 이름[6], 문장 사용 패턴[7], 어휘적 및 문법적 특성[8]을 분석하여 성별을 추론하는 데 사용된다. 이러한 방법들은 명확한 성별 지시자를 가진 이름에 효과적이나, 중성적 이름이나 다양한 문화적 맥락에서 한계를 드러낸다. 또한, 소셜 미디어 메시지 데이터를 이용한 성별 예측 연구[9]는 비정형 데이터에서 높은 예측력을 보이나, 특정 사용자 그룹에 한정될 수 있다는 한계를 지닌다.

2.2. NLP 에서 성별 편향 완화 연구에 관한 연구

성별 편향 완화는 NLP 모델의 공정성과 정확성을 높이기 위한 중요한 연구 주제이다. 연구자들은 단어 임베딩에서 성별 편향을 감지하고 제거하는 방법[10], 성별 중립적 임베딩을 학습하는 시도[11], 관계 추출[12] 및 자연어 추론 분야[13]에서 성별 편향을 평가

하고 완화하는 기법 등을 개발해 왔다. 또한, 코퍼스 분야의 연구는 성별 정보를 문장에 추가하고 기존 모델을 재학습시켜 편향을 완화하는 방식으로 접근한다[14]. 이러한 접근법은 텍스트 코퍼스 내의 성별 관련 표현을 재조정하여 기계 학습 모델이 성별 정보를 더 균형 잡힌 방식으로 처리하도록 한다.

2.3. 제로샷 분류 기법

제로샷 분류(Zero-shot Classification)는 기존에 학습한 데이터와 유사성을 바탕으로, 훈련 중에 보지 못한 새로운 클래스를 분류할 수 있는 능력을 갖춘 기법이다[15]. 이 방법은 주로 사전 훈련된 언어 모델을 기반으로 하여 텍스트의 의미적 특징을 파악하고, 이를 활용하여 분류 작업을 수행한다[16]. 제로샷 분류의 핵심 기능은 모델이 학습 데이터셋에 직접 나타나지 않은 새로운 클래스를 예측할 수 있게 하는 능력이며, 이는 특히 라벨링이 어려운 데이터셋에서도 뛰어난 성능을 발휘한다.

3. 성별 편향이 완화된 성별 예측

본 연구에서는 성별 예측의 정확성을 높이고 성별 편향을 최소화한 성별 예측 방법을 제시한다. 본 연구에서는 이를 위해 성별 클래스를 확장하고 데이터 구조를 변형한 데이터를 제작하여 사용하였다. 또한, 기존의 제로샷 분류 기법에 성별 클래스를 세분화하고, 더욱 포괄적인 성별 카테고리를 인식할 수 있도록 기존의 모델을 변형하여 이를 성별 예측에 적용하였다.

3.1. 성별 편향 완화 데이터

성별 클래스 확장의 경우 기존의 '여성'과 '남성'의 이분법적인 성별 분류 방식을 넘어, '여성', '남성', 그리고 '여성 또는 남성'으로 세분화한다. 이는 성별을 더욱 정밀하게 예측할 수 있게 하여, 문맥상 성별이 명확하지 않은 비특정 성별의 경우에도 유연하게 성별을 예측할 수 있게 한다. 또한, AI Hub 에서 수집한 문장 데이터에서 단일 인물을 추출하고, 인물이 의미하는 성별 데이터를 추가하여 데이터 구조를 변형하였다. <표 1>은 변형된 데이터 구조를 보여준다. <표 1>과 같이, 데이터는 '문장'-'문장 내 등장하는 단일 인물'-'인물이 해당 문장에서 의미하는 성별'로 구성되어 성별 예측의 정확도를 높이는 데 중요한 역할을 한다. 이렇게 생성된 성별 편향을 최소화한 한국어 데이터를 중심으로 하여 제로샷 분류 기법을 활용한 성별 예측 방법을 제안한다. 이와 함께, 제로샷 분류 모델의 제한된 환경에서의 범용성을 검증하기 위해 소량의 영어, 중국어 데이터를 제작하여 사용하였다. <표 2>, <표 3>, <표 4>는 각각 한국어, 영어, 중국어 데이터 샘플을 보여준다. 이렇게 생성된 한국어 데이터는 Base 데이터 약 12,440 개와 이를 약간 증강한 30,053 개이며, 영어, 중국어 데이터는 각각 500 개씩 생성되었다.

3.2. 제로샷 분류 기반 성별 예측 모델

변형된 접근법은 3.1 장에서 언급한 데이터를 사용해 추가적인 특성을 학습하여 성별 클래스를 더욱 세분화하고, 모델이 성별 예측 시 더 다양하고 포괄적인 성별 카테고리를 인식할 수 있도록 확장한다. 이러한 확장은 특히 다양한 문화적 및 언어적 배경을 가진 데이터셋에서 성별을 더 정확하게 예측할 수 있게 한다.

<표 1> 변형된 데이터 구조

	문장	인물	정답 성별
내용	특정 언어의 한 문장	문장 내 등장하는 단일 인물	문장 내에서 인물이 의미하는 성별
예시	'철수는 졸업을 앞두고 있다.	'철수'	'여성 또는 남성'

<표 2> 한국어 데이터 샘플

문장	인물	정답 성별
언니의 이름은 사드나이고 **저**보다 4 살 많습니다.	저	'여성'
AAA 의 단편 공포에는 자신의 친구 **실린**의 아내 마리에게 매혹되는 주인공 이자 1 인칭 화자인 인물이 등장한다.	실린	'남성'
나는 그녀와 일요일 날 만날 예정입니다.	나	'여성 또는 남성'

<표 3> 영어 데이터 샘플

문장	인물	정답 성별
B is the wife of a patient discharged from Ward 81 who was diagnosed as confirmed on the 8th.	B	'female'
A sweet phone call between Sung Yu-ri, a member of the group Fin.K.L, and her husband Ahn **Sung-hyun** has been released.	Sung-hyun	'male'
Mental health specialists include one mental health **nurse** and two mental health social workers.	nurse	'female or male'

<표 4> 중국어 데이터 샘플

문장	인물	정답 성별
我的老公就是上班族.	我	'女性'
送给我家人和女朋友的礼物.	我	'男性'
我在棒球场的西边看到你了， 我们去那里好吗？	你	'女性或男性

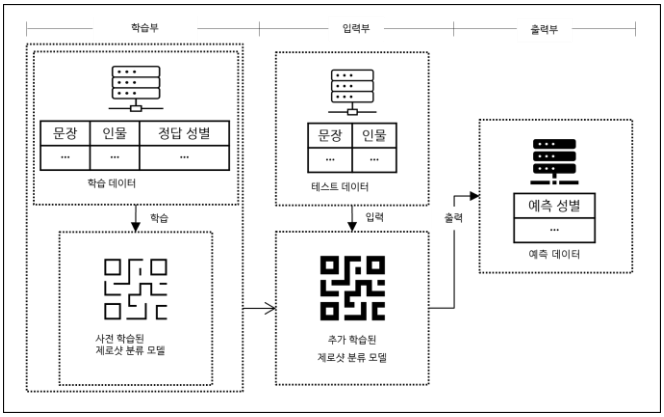
(그림 1)은 제로샷 분류 기반 성별 예측 모델의 아키텍처를 나타낸다. 제로샷 분류 모델은 Hugging Face 에서 제공하는 대량의 다언어 데이터를 사전 학습한 모델인 ZS1(MoritzLaurer/mDeBERTa-v3-base-mnli-xnli), ZS2(MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil)을 기반으로 하여, 3 장의 성별 편향 최소화 데이터를 활용하여 추가로 학습을 진행한다. 모델은 학습부, 입력부, 출력부로 구성되어 있으며, 학습부에서는 성별 편향을 완화하고 예측 정확성을 높이기 위해 추가 학습을 진행한다. 여기서 사용된 데이터는 '문장'과 '문장 내 등장하는 단일 인물'을 이어 붙인 데이터와 '정답 성별' 데이터로 구성되어 있다. 입력부에서는 성별 예측을 위한 '문장'과 '문장 내 등장하는 단일 인물' 데이터를 각각 입력으로 받아들여, 추가 학습된 제로샷 분류 모델은 입력에 대한 성별 예측을 수행한다. 출력부에서는 최종적인 성별 예측 결과를 제공한다.

4. 실험

4.1. 제로샷 분류 기반 성별 예측 실험

본 실험은 제로샷 분류 기반 성별 예측 모델의 성능을 지도 학습 모델과 비교하여 성별 예측의 정확성

과 효과를 검증한다. 제로샷 분류 모델은 3 장에서 언



(그림 1) 제로샷 분류 기반 성별 예측 모델 아키텍처

급한 두 가지 사전 학습된 모델을 기반으로 한국어 Base 데이터와 증강 데이터를 추가 학습한 후 성별 예측을 수행하였다. 성별 예측의 정확도를 평가하기 위해, 결정 트리, k-인접 이웃, 그래디언트 부스팅 머신, 서포트 벡터 머신 네 가지 지도 학습 모델을 대조군으로 사용하였다. <표 5>는 본 실험의 결과를 나타낸다. 실험 결과, 제로샷 분류 모델(ZS1, ZS2)은 지도 학습 모델(DT, KNN, GBM, SVM)보다 높은 정확도를 보였다. ZS1 모델의 경우 Base 데이터에서 89.01%의 정확도를 달성했으며, 이는 지도 학습 모델에서 가장 높은 성능을 보인 SVM 의 77.41%보다 약 11.6% 더 높았다. 증강 데이터에서도 ZS1 은 88.87%의 정확도로, SVM 의 77.65%보다 11.22% 더 높은 결과를 보였다. Macro F1 점수에서도 제로샷 모델이 지도 학습 모델보다 우수한 성능을 나타냈다.

<표 5> 제로샷 분류 모델 및
지도 학습 분류 모델 기반 성별 예측 결과

모델	Base 데이터		증강 데이터	
	Accuracy	Macro F1	Accuracy	Macro F1
ZS1	89.01	88.47	88.87	88.74
ZS2	88.76	88.21	87.93	87.76
DT	72.57	71.67	70.69	70.41
KNN	66.41	66.39	70.41	70.05
GBM	73.06	70.00	70.84	70.69
SVM	77.41	75.78	77.65	77.70

4.2. 모델 범용성 검증 실험

본 실험은 제로샷 분류 기반 성별 예측 모델의 범용성을 평가하기 위해 한국어, 영어, 중국어 데이터를 사용하여 제로샷 분류 모델(ZS1, ZS2)에 추가 학습시켜 성별 예측을 수행하였다. <표 6>은 본 실험의 결과를 나타낸다. 실험 결과, 제로샷 분류 모델은 한국어, 영어, 중국어 데이터에서 높은 정확도를 유지했다. 특히, 중국어 데이터에서는 최대 88%의 높은 정확도를 보여주었으며, 이는 언어의 특성과 데이터의 단순성이 모델 성능에 긍정적인 영향을 미친 것으로 분석된다. 실험을 통해, 제로샷 분류 모델은 제한된 데이터 양에서도 일관된 성능을 보여주어, 다언어의 소량 데이터에서도 효과적인 성별 예측이 가능함을 입증했다.

<표 6> 다언어 및 소량의 데이터 환경에서의
제로샷 분류 모델 성별 예측 결과

모델	한국어 데이터		영어 데이터		중국어 데이터	
	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
ZS1	80	80.83	80	78.75	86	83.34
ZS2	78	77.47	78	76.68	88	85.73

4.3. 실험 분석 및 한계

본 실험은 제로샷 분류 기반 성별 예측 모델의 성능과 범용성을 평가하는 데 주안점을 두었다. 실험은 지도 학습 분류 모델과의 성능 비교를 통해 제로샷 분류 모델의 정확성과 효율성을 측정하고, 한국어, 영어, 중국어 데이터를 활용하여 제로샷 분류 모델이 다양한 언어 환경에서도 효과적으로 성별을 예측할 수 있는지 검증하였다. 실험 결과, 제로샷 분류 모델이 지도 학습 모델에 비해 높은 성능을 보이며, 소량의 다양한 언어 데이터 환경에서도 일관된 예측 성능을 보여주었다.

제로샷 분류 기법의 이러한 능력은 특히, 학습 데이터양이 제한된 환경에서 중요한 의미가 있다. 실험에서는 한국어, 영어, 중국어 데이터 각각 500 개씩만을 사용했음에도 불구하고, 모델은 비교적 높은 정확도와 Macro F1 점수를 달성했다. 이 결과는 제로샷 분류 모델이 제한된 데이터 환경에서도 효과적으로 예측을 수행할 수 있음을 입증하며, 데이터 부족 문제를 겪고 있는 비영어권 언어 연구에 있어 중요한 해결책을 제시한다.

실험 결과를 통해 성별 클래스 불균형과 클래스 중첩 문제가 예측 성능에 영향을 주는 것으로 확인되었다. 성별 클래스 불균형은 특정 성별 클래스의 데이터양이 다른 클래스에 비해 상대적으로 많거나 적을 때 발생하는 문제이다. 또한, '여성' 클래스를 '남성' 클래스로, 또는 그 반대로 잘못 분류된 결과도 확인되는데, 이는 사전 학습 데이터에 내재된 성별 편향의 영향과 성별 클래스 불균형이 모델의 예측 성능에 영향을 준 것으로 해석된다. 클래스 중첩 문제의 경우, '여성 또는 남성' 클래스가 '여성'이나 '남성' 클래스로, 또는 그 반대로 잘못 분류된 경우를 의미하는데, 이는 모델이 '여성 또는 남성' 클래스를 충분히 이해하지 못했음을 시사한다. 이에 대한 대응으로, 성별 클래스의 데이터 비율을 조정하여 데이터셋 내에

서 균형을 맞추고, 다양한 성별 클래스 표현 방식을 탐색할 필요가 있다. 이를 통해 모델이 각 성별 클래스를 보다 효과적으로 인식하여, 결과적으로 성별 예측의 정확성을 높일 수 있을 것으로 예상된다.

5. 결론

본 연구는 자연어 처리 분야에서 성별 예측과 성별 편향의 완화를 목표로 진행되었다. 기존의 연구들이 주로 영어 데이터에 집중됐지만, 본 연구는 한국어를 포함한 다양한 언어 데이터의 활용에 중점을 두었다. 이 연구는 더 포괄적이고 정확한 성별 예측을 가능하게 하는 제로샷 분류 기반 모델을 제안하였으며, 이 모델의 성능을 다양한 언어 및 제한된 데이터 환경에서 검증하였다.

연구 결과, 제안한 모델은 제한된 데이터 환경에서도 높은 정확도를 유지하며, 다양한 언어 데이터에서 성별을 효과적으로 예측할 수 있음을 입증하였다. 이는 모델이 성별 편향을 효과적으로 완화하고, 성별 클래스를 확장하여 보다 포괄적으로 성별을 인식할 수 있음을 보여준다.

본 연구의 접근법은 자연어 처리를 넘어 의료, 보안 등의 다양한 분야에서 응용 가능성을 제시한다. 제로샷 분류 기법을 활용한 성별 예측 모델은 사용자의 성별을 보다 정확하고 포괄적으로 파악하여 개인화된 서비스 제공에 기여할 수 있다. 향후에는 데이터 제작 자동화와 모델의 성능을 더욱 향상하는 연구를 수행할 예정이다.

Acknowledgement

이 연구는 2022 학년도 이화여자대학교 교내연구비 지원과 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.RS-2022-00155966, 인공지능융합혁신인재양성(이화여자대학교))에 의한 연구임.

참고문헌

[1] Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., ... & Roth, D. “Recent advances in natural language processing via large pre-trained language models: A survey.” ACM Computing Surveys, 56(2), 1-40 (2023).

[2] Prates, M. O., Avelar, P. H., & Lamb, L. C. “Assessing gender bias in machine translation: a case study with google translate.” Neural Computing and Applications, 32, 6363-6381. (2020).

[3] Levy, S., Lazar, K., & Stanovsky, G. “Collecting a large-scale gender bias dataset for coreference resolution and machine translation.” arXiv preprint arXiv:2109.03858. (2021).

[4] Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. “Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods.” Big data and cognitive computing, 7(1), 15 (2023).

[5] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Cruz, G. O. R., Peixoto, R. M., Guimarães, G. A. D. S., ... & Nascimento, E. G. S. “Bias and unfairness in machine learning models: a systematic literature review.” arXiv preprint arXiv:2202.08176. (2022).

[6] Wais, K. “Gender prediction methods based on first names with genderizeR.” R J., 8(1), 17. (2016).

[7] Veronika, C. V. A. S. T., Illés, S. Z., & Dahiya, S. “Gender prediction of the European school’s teachers using machine learning: Preliminary results.” In Proceeding of 8th IEEE International Advance Computing Conference, India, (2018, December) (pp. 213-220).

[8] van der Goot, R., Ljubešić, N., Matroos, I., Nissim, M., & Plank, B. “Bleaching text: Abstract features for cross-lingual gender prediction.” arXiv preprint arXiv:1805.03122. (2018).

[9] Abdallah, E. E., Alzghoul, J. R., & Alzghool, M. “Age and gender prediction in open domain text.” Procedia Computer Science, 170, 563-570. (2020).

[10] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings.” Advances in neural information processing systems, 29. (2016).

[11] Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. “Learning gender neutral word embeddings.” arXiv preprint arXiv:1809.01496. (2018).

[12] Gaut, A., Sun, T., Tang, S., Huang, Y., Qian, J., ElSherief, M., ... & Wang, W. Y. “Towards understanding gender bias in relation extraction.” arXiv preprint arXiv:1911.03642. (2019).

[13] Rudinger, R., May, C., & Van Durme, B. “Social bias in elicited natural language inferences.” In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Spain (2017, April) (pp. 74-79).

[14] Vanmassenhove, E., Hardmeier, C., & Way, A. “Getting gender right in neural machine translation.” arXiv preprint arXiv:1909.05088. (2019).

[15] Yin, W., Hay, J., & Roth, D. “Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.” arXiv preprint arXiv:1909.00161. (2019).

[16] Puri, R., & Catanzaro, B. “Zero-shot text classification with generative language models.” arXiv preprint arXiv:1912.10165. (2019).