Journal of the Korean Institute of Industrial Engineers
Vol. 48, No. 6, pp. 636-643, December 2022.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2022.48.6.636

악성 댓글에 대한 한국어 혐오표현 및 편견 탐지 분류 모형 결과 분석 및 개선방안 연구

이세영[1] · 박새롬[2†]

[1]성신여자대학교 미래융합기술공학과 / [2]성신여자대학교 융합보안공학과

Analyzing the Classification Results for Korean Hatespeech and Bias Detection Models in Malicious Comment Dataset

Seyoung Lee[1] · Saerom Park[2]

[1]Department of Future Convergence Technology Engineering, Sungshin Women's University
[2]Department of Convergence Security Engineering, Sungshin Women's University

With the development of Internet communication technology, opinions on various issues can be freely expressed on the Internet. However, some people have abused their freedom of expression, causing psychological harm by writing comments expressing their hatred towards others. In order to address this problem, research on automatic detection of malicious comments using machine learning models has been actively conducted. In this study, we constructed the detection models for hate speech and bias to classify KOCO (KOrean hate COmments) dataset using popular language classification models such as logistic regression with term frequency-inverse document frequency, KoBERT, KoELECTRA, KcELECTRA and KoGPT2 models. Through the experiments, we demonstrated that sentence length, reflection of context information, and mis-labeled data highly affected the classification performance of most models. As a result, we presented considerations for automatic detection of malicious comments and directions for constructing the comment dataset to improve the detection models in future research.

Keywords: Korean Hatespeech Classification, Bias Classification, Malicious Comments

1. 서 론

현대 사회에서는 초고속 인터넷이 발달하고 스마트폰이 보급으로 인해 시간과 장소에 구애받지 않고 누구나 인터넷을 사용한다. 모바일 기기로 손쉽게 인터넷을 사용할 수 있게 되면서 유튜브, 인스타그램 등의 SNS가 발달하였고, 다양한 SNS를 통해 각종 이슈에 대한 자신의 의견을 댓글로 표현할 수 있게 되었다. 하지만, 이를 악용하여 상대방에 대한 비방, 비난, 조롱, 혐오 등을 표현하는 댓글을 작성하여 상대방에게 정신적인 피해를 입히기도 한다(Moon *et al.*, 2022). 이에 따라 악성댓글로 인한 피해를 막기 위하여 인공지능을 활용한 악성댓글 분류 모델에 대한 연구가 활발히 이루어지고 있다.

악성댓글 분류 모델을 학습시킬 수 있도록 레이블링 되어 공개된 한국어 데이터셋은 KOCO(KOrean COmments) dataset(Moon *et al.*, 2022), Unsmile dataset(Kang *et al.*, 2022) 등이 있다. KOCO dataset은 Kaggle competition으로 진행되어 다양한 모델의 성능을 리더보드를 통해 비교할 수 있다. 혐오 표현 분류에서 가장 높은 성능을 보인 모델도 f1-score가 0.68로 실제 온라인 서비스에 적용하기에 부족한 성능을 보인다. 따라서 본 연구에서는 인공지능을 활용한 혐오 표현 탐지의 성능

이 논문은 2022년도 교육부 및 산업통상자원부의 재원으로 한국연구재단 기초연구사업(과제번호: NRF-2022R1F1A1065171)과 한국산업기술진흥원 산업혁신인재성장지원사업(P0008703)의 지원을 받아 수행된 연구임
† 연락저자 : 박새롬 교수, 02844 서울특별시 성북구 보문로 34다길 2 프라임관 503호, Tel : 02-920-7599, E-mail :psr6275@sungshin.ac.kr
2022년 6월 13일 접수; 2022년 8월 7일 수정본 접수; 2022년 8월 29일 게재 확정.

개선을 위해 악성 댓글 데이터의 특징을 살펴보고 혐오 및 편견 분류 문제에 대한 인공지능 모형의 예측 결과를 비교하고자 한다. 이를 위해 기계학습 모형인 Logistic Regression(LR)과 딥러닝 모형인 KoBERT(SKTBrain. 2019), KoGPT2(SKT-AI. 2021), KoELECTRA(Park. 2021), KcELECTRA(Jumbum. 2021)를 사용하였다. KOCO dataset의 혐오 댓글과 편견 분류 문제를 각 모델에 학습시켜 모델별 성능을 비교하고, 각 문제와 관련된 데이터의 특성을 분석한다. 또한, 예측된 결과에 대한 설명을 제공하는 LIME (Local Interpretable Model-agnostic Explanation) (Ribeiro *et al.*, 2016) 알고리즘을 이용하여 분류 작업에서 영향을 미친 요인을 분석하여 향후 데이터 수집 및 연구에서 개선되어야 할 부분을 제시하고자 한다.

2. 관련 연구

인터넷의 사용이 전 세계적으로 보편화되면서 악의적 댓글은 기하급수적으로 늘어났다. 인터넷 사용에 대한 시공간의 제약이 없어지고, 사용자 수가 급격하게 증가하면서 댓글의 업로드 속도 또한 매우 증가하였다. 따라서 사이트 관리자가 직접 댓글을 검수하여 악의적인 댓글을 차단하는 것은 거의 불가능해지면서 인공지능을 활용하여 악성 댓글을 관리하고자 하는 연구가 활발해지고 있다.

감성 분석과 Support Vector Machine(SVM)을 활용하여 악성 댓글을 탐지하려는 연구(Jinju et al, 2016) 에서는 단어의 악의성을 0~1의 수치로 표현하고, 단어의 악의성 수치를 활용하여 SVM 모델로 악성 댓글을 탐지하려는 시도를 하였다. 학습된 모델의 성능은 재현율이 87.8%로 이전 연구들보다 좋은 성능을 보였지만, 주기적으로 단어 사전과 단어의 악의성 수치를 업데이트해야 하고, 문장 내 단어의 쓰임에 따른 문맥 정보를 파악할 수 없다는 한계가 있었다.

KoELECTRA 모델에 KOCO dataset과 심심이 나쁜말 데이터셋을 학습시켜 분류 성능을 확인 및 비교한 연구(Shin. *et al.*, 2021)에서는 한국어 악성댓글 데이터셋과 심심이 나쁜말 데이터셋에서 각각 f1-score가 0.63, 0.66 이 나오는 것에 대해 모델 성능에 대한 정성분석을 실시하였다. 모델 성능이 좋지 못한 이유로 댓글이나 챗봇 데이터의 경우 비정형 데이터가 많고, 맥락적 정보를 잘 파악하지 못한 경우가 있었으며, 데이터 레이블링이 잘못된 경우 때문이라고 주장하였다. 심심이 나쁜말 데이터셋을 주로 분석하였고, 혐오 표현 및 편견 발언 분류 레이블이 존재하는 KOCO dataset에 대한 정성적 분석이 부족하였다. 본 연구에서는 KOCO dataset만을 분석하며, 혐오표현 분류와 편견 발언 분류에 대한 성능 차이에 대한 분석도 추가적으로 진행한다. 또한, BERT기반의 KoBERT, KcELECTRA 모델과 GPT기반의 KoGPT2 모델을 추가로 실험하여 다양한 딥러닝 모형의 성능을 비교한다.

형태소 분석을 통한 악성댓글 필터링 방안 연구(Yeram. *et al.*, 2021)에서는 온라인 댓글이 비방, 욕설, 비하 등 사이버 언어폭력에 해당하는 발언인지를 의미하는 수치를 댓글 충격량으로 정의하고 형태소분석을 통해 댓글 충격량을 분석하여 이를 바탕으로 악성 댓글 필터링 방안을 제안하였다. 네이버 클린봇과 비교하였을 때 제안하는 방법을 통해 f1-score가 47.66% 향상되었지만, 비속어 및 존대어 형태소 분석을 기초로 하고 있어 데이터베이스의 최신화에 의존적이기 때문에, 최신화가 되지 않은 경우, 신조어 욕설에 대한 성능이 떨어졌다. 또한, 문맥 정보를 활용하지 못하는 한계점이 존재하여 단어의 원래 의미로 사용되지 않고 문맥상 비속어처럼 사용된 경우에는 단순히 형태소만으로 분류하기 어려운 한계가 있었다. 따라서, 본 연구에서는 모델별 분류 결과를 통해 모델들이 문맥 정보를 반영하고 있는지, 욕설, 비속어, 비방, 비하 등 악의성을 내포한 단어를 탐지하고 있는지 등에 대한 정성적 분석을 실시하고자 한다.

3. 실험 방법

3.1 데이터셋

<Table 1>의 KOCO dataset은 인터넷 연예 기사에 대한 한국어 댓글 데이터셋이다. KOCO dataset은 특정 집단에 대한 편견이나 차별에 따라 gender bias, others bias, none bias 클래스로 분류하고, 혐오 정도에 따라 hate, offensive, none 클래스로 분류하여 multi-label을 가지는 약 1만 개의 댓글 데이터이다. 데이터셋을 구축한 연구팀에서는 데이터셋에 대해 char-CNN, BiLSTM, KoBERT 모델로 분류한 성능을 제시하였고, Kaggle Competition을 통해 데이터셋의 분류 성능을 높이기 위한 방법을 모색하고 있다.

3.2 데이터 전처리

데이터 전처리 과정은 <Table 2>에서와 같이 각 모델에 적합한 방식으로 진행되었다. LR 모델의 입력으로 Word2vec을

Table 1. The Number of Data According to Label Information in Koco Dataset

	Hatespeech			Bias			total
	offensive	hate	none	gender	others	none	
train	2499	1911	3486	1232	1516	5148	7896
test	189	122	160	67	62	342	471

Table 2. Pre-processing Configuration

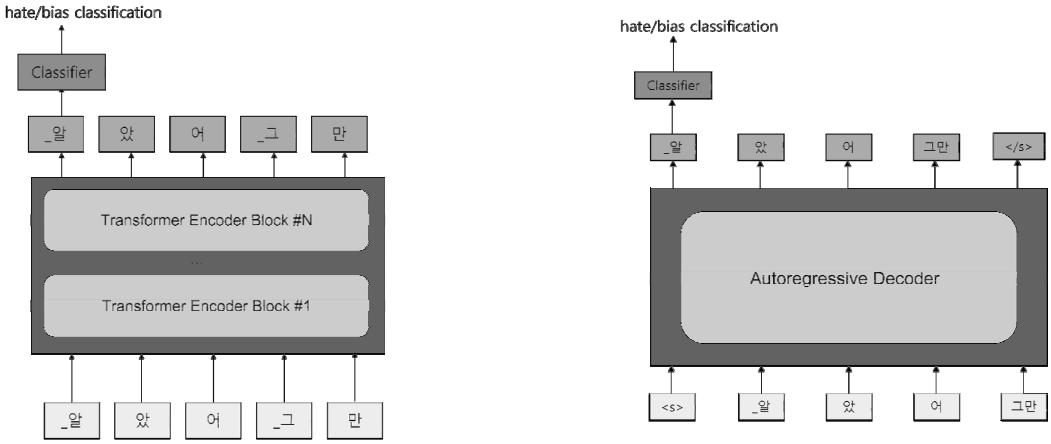
	Tokenizer	Vocab size	max length	OOV token
LR	soynlp	100000	-	-
KoBERT	kobert tokenizer	8002	64	UNK
KoELECTRA	electra tokenizer	32200	64	UNK
KoGPT2	GPT2Tokenizer	51200	64	UNK
KcELECTRA	electra tokenizer	50135	64	UNK

한국어에 활용한 kor2vec(Junsung. 2018) 을 사용하였을 때 분류 성능이 좋지 않았기 때문에 tf-idf 방법론을 사용하여 전처리를 진행하였다. 온라인 댓글 특성상 'ㅋㅋㅋ', 'ㅎㅎㅎ'와 같은 자음 반복이 많이 발견되기 때문에 의미 없는 반복 문자열은 'ㅋ'이나 'ㅎ'과 같이 반복 문자열 중 한 글자만 남기고 제거하였고, 반복 문자열 제거 후 8자 미만의 데이터는 삭제하였다. 정리된 자연어 데이터를 기계학습 모델에 사용하기 위해 tfidf-vectorizer를 사용하여 벡터화 하였고, n-gram기법을 적용하여 1-3단어로 묶이도록 하였다. tfidf-vectorizer의 단어사전 크기는 100000으로 설정하였고, 문장의 최대 길이는 설정하지 않았다. 단어사전 외의 단어는 토큰화하지 않고 삭제하였다.

KoBERT를 비롯한 딥러닝 모형들은 사전학습과정에서 미리 학습된 토크나이저를 사용하여 전처리 과정을 진행하였다. 토큰화된 문장의 최대 길이는 64로 설정하였다. 단어사전 외의 단어는 UNK 토큰으로 치환하였다. 각 딥러닝 모형들은 혐오 표현 분류 모델과 편견 발언 분류 모델을 각각 구축하여 사전학습 모델을 KOCO datset으로 재학습시킨 후 예측을 수행하였다.

3.3 분류 모형

본 연구에서는 예측 성능 및 결과의 비교를 위해서 머신러닝 알고리즘인 LR모델과 딥러닝 모델인 KoBERT(SKTBrain. 2019), KoGPT2(SKT-AI. 2021), KoELECTRA(Park. 2021), KcELECTRA(Jumbum. 2021) 모형을 사용하였다. LR모델은 입력 데이터가 어떤 클래스에 속할 확률을 0~1 사이의 값으로 예측하고 그 확률에 따라 가능성이 더 높은 클래스에 속하는 것으로 분류하는 선형 분류 모형이다. <Figure 1(a)>의 KoBERT는 구글이 공개한 BERT 모델에 한국어 대용량 데이터셋을 학습시킨 사전학습모델로, 다양한 한국어 자연어처리 작업에서 우수한 성능을 보이는 모델이다. BERT 모델은 입력 문장의 일부를 MASK 토큰으로 치환하고, 이를 다시 원본 Token으로 복원하는 방식으로 사전학습을 수행한다. ELECTRA 모델은 BERT 기반의 모델로, 입력 문장의 일부를 MASK 토큰이 아닌 가짜 토큰으로 치환하고, 이를 다시 원본 토큰으로 복원하는 사전학습과정을 수행하여 성능을 높인 모델이다. KoELECRTA, KcELECTRA 모델은 ELECTRA 모델에 한국어 대용량 데이터셋을 학습시킨 딥러닝 모델이라는 공통점이 있지만, 학습에 사용한 데이터셋이 달라 서로 다른 분야에서 좋은 성능을 보인다. KoELECTRA 모델은 한국어 위키백과, 뉴스데이터, 모두의 말뭉치 등의 한국어 대용량 데이터셋을 학습시켜 정형화된 데이터를 처리할 때 좋은 성능을 보이고, KcELECTRA는 한국어 뉴스 기사 댓글 데이터셋을 학습시켜 댓글 데이터와 같은 비정형 데이터를 처리할 때 비교적 좋은 성능을 보인다. KoGPT2 모델의 구조는 <Figure 1(b)>와 같다. GPT모델은 Encoder 구조를 사용한 BERT와 달리 Decoder만을 사용하여 학습한다. GPT모델은 주어진 문장의 다음단어를 예측하는 작업에 최적화된 모델이지만, 본 연구에서는 마지막 레이어를 classifier로 구성하여 혐오 및 편견 문장 분류를 학습시켰다.



(a) KoBERT architecture (b) KoGPT2 architecture

Figure 1. Deep Learning Model Architecture

3.4 모델 해석 알고리즘

LIME 알고리즘은 블랙박스 모델에 대해 해석 가능한 대리 모델(surrogate model)을 통해 모델을 해석하고자 하는 알고리즘이다. LIME에서는 입력 x가 주어졌을 때, 해석 가능한 모델들의 집합 G에서 신경망 모델 f와 가장 근사한 결과를 예측하는 해석 가능한 모델 g∈G를 찾아 신경망 모델에 대한 해석을 제공한다. 본 연구에서는 입력 데이터에서 어떤 부분이 모델의 예측에 영향을 미쳤는지 분석하고자 한다. LIME 모형을 통해서 입력 데이터의 가중치를 확인하여 모델의 추론 과정에서 각 토큰이 어떤 영향을 미치는지 확인하였다.

4. 실험 결과 및 분석

4.1 혐오 발언 분류

다음 <Table 3>은 혐오 표현 분류 작업에 대한 LR모델, KoBERT, KoELECTRA, KoGPT2, KcELECTRA의 성능을 나타낸다. 주어진 모델 중에서는 KcELECTRA의 성능이 정확도 및 F1-score 0.68로 가장 좋았지만, 좋은 성능이라고 보기 어렵다. 특히, LR모델과 KoBERT을 비교하였을 때, KoBERT 모형은 다른 자연어처리 작업들에서 우수한 성능을 보이는 딥러닝 모델임에도 불구하고 두 모델 사이의 성능은 거의 유사하였다. 본 연구에서는 LIME을 통해 모델의 예측 과정을 살펴보고 모델의 예측 성능에 영향을 끼치는 요인을 분석하였다.

(1) 문맥 정보의 반영

본 연구에서 사용한 기계학습 모형 및 딥러닝 모형에서는 단어사전에 등록되지 않은 신조어나 오탈자는 삭제되거나 [UNK]토큰으로 치환되었다. 각 단어의 의미보다 문장의 문맥 정보를 학습하기 위해 LR모델에서는 n-gram 방법을 사용하였다. KoBERT와 같은 딥러닝 모형에서는 OOV 문제를 줄이기 위해 BPE(Byte Pair Encoding) 방식을 변형한 WordPiece 방식으로 토크나이징을 수행하였기 때문에 한글로 이루어진 신조어나 오탈자가 [UNK]토큰으로 치환되는 경우는 매우 적었고, 한자와 이모티콘이 대부분 [UNK]토큰으로 치환되었다. 또한, KoBERT를 비롯한 딥러닝 모형들은 attention 기반의 딥러닝 모형으로, 문장 내 단어 간의 관계를 학습한다.

기계학습 모형인 LR모델과 딥러닝 모형 중 성능이 가장 좋았던 KcELECTRA 모형에서 같은 단어가 다른 의미로 쓰였을 때 문맥적인 정보를 파악하여 서로 다른 의미의 단어로 확인할 수 있는지 LIME을 통해 확인하였다. <Figure 2>는 LR모델에서 '일본'이라는 단어가 등장한 서로 다른 두 개의 문장에서 label에 미치는 영향이 다른 것을 보여주는 LIME 분석 결과다. '일본'이라는 단어는 욕설이나 혐오적인 발언이 아님에도 offensive가 아니라 hate에 영향을 주고 있는 것을 확인할 수 있었고, 두 문장에서 '일본'이 hate label에 주는 영향의 차이가 적은 것을 확인할 수 있었다. <Figure 3>은 KcELECTRA 모델에서 '일본'이라는 단어가 서로 다른 두 개의 문장에서 offensive label에 미치는 영향이 다른 것을 보여주는 LIME 분석 결과다. 각 문장에서 KcELECTRA 모델은 attention 기반의 딥러닝 모

Table 3. Perfomance of hatespeech and bias classification in KOCO Dataset

	Hatespeech		Bias	
	Accuracy	F1-score	Accuracy	F1-score
LR	0.62	0.61	0.73	0.66
KoBERT	0.62	0.61	0.80	0.69
KoELECTRA	0.65	0.64	**0.83**	**0.74**
KoGPT2	0.57	0.56	0.75	0.64
KcELECTRA	**0.68**	**0.68**	0.80	0.72

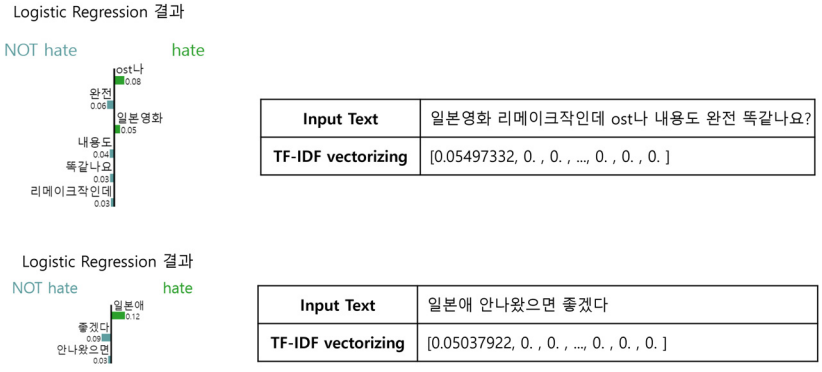
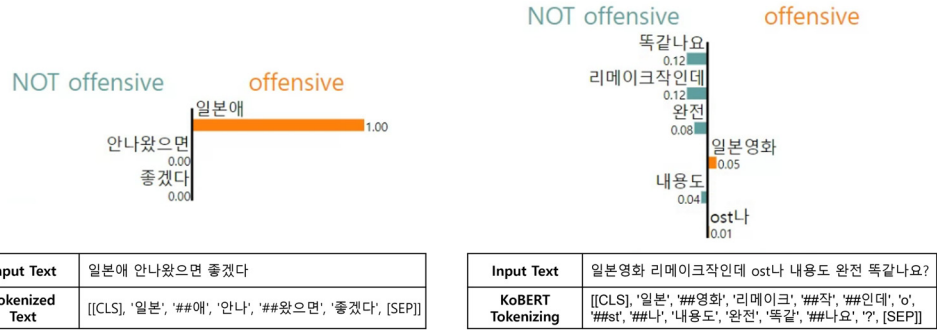


Figure 2. Context Analysis with Logistic Regression



Input Text	일본애 안나왔으면 좋겠다
Tokenized Text	[[CLS], '일본', '##애', '안나', '##왔으면', '좋겠다', [SEP]]

Input Text	일본영화 리메이크작인데 ost나 내용도 완전 똑같나요?
KoBERT Tokenizing	[[CLS], '일본', '##영화', '리메이크', '##작', '##인데', 'o', '##st', '##나', '내용도', '완전', '똑같', '##나요', '?', [SEP]]

Figure 3. Context Analysis with KcELECTRA

델이기 때문에 같은 단어라도 각 문장의 문맥에 따라 영향을 준 label이 다른 것을 확인할 수 있었다.

따라서 한국어 혐오 표현 분류를 위해서는 문맥을 잘 학습할 수 있는 딥러닝 모형이 필요하고, 특정 단어가 다양한 문맥에서 사용된 예제 문장들을 더 추가하여 문맥에 따른 단어의 다양한 쓰임새를 학습시켜야한다.

(2) 잘못 레이블링된 데이터

<Table 3>에서 혐오 표현 분류 모형의 경우, 좋지 못한 성능을 보였기 때문에 주어진 데이터로 학습된 모형에 대한 LIME 분석을 통해 예측 결과에 영향을 준 요인들에 대해서 살펴보았다. 하지만, 해당 실험에 사용한 데이터셋이 크라우드소싱 방법으로 레이블링하여 구축되었으므로 모형 학습에 사용된 데이터 자체의 문제점이 있는지를 파악하기 위해서 데이터셋을 직접 살펴보고 데이터 태깅 가이드라인과 다르게 태깅이 되어있는 레이블링 오류를 확인하고자 하였다.

<Table 4>는 혐오 표현 분류 문제에서의 레이블이 잘못 표기된 각 클래스의 대표적인 문장들이다. 잘못된 첫 번째 문장에서는 10+8이라는 발음상 욕설이 되는 분명한 욕설이 있음에도 불구하고 none, 즉 정상 데이터로 레이블링 되었다. 또한 두 번째 문장은 단순히 바람을 말하는 정상 데이터로 볼 수 있음에도 offensive, 공격적인 어조를 가진 댓글로 분류되었다. 마지막으로 세 번째 문장은 hate로 레이블링이 되었는데, 작가의 성(姓)을 말하는 'ㅈ' 을 어노테이터들이 욕설로 해석하였을 것으로 예측할 수 있다.

본 연구에서 사용한 데이터셋의 경우에는 데이터 레이블링 작업을 크라우드소싱으로 진행하였기 때문에 데이터 레이블링의 오류가 존재할 수 있다. 하지만, 데이터 레이블링의 오류로 인해 인공지능 모델의 성능이 매우 저하될 수 있으므로, 이를 보완하기 위해 레이블 노이즈를 고려한 다양한 기법의 적용이 필요하다. 예를 들면, 더 많은 데이터를 레이블링하거나 크라우드소싱의 어노테이터들의 합의 지수가 높은 데이터들을 우선적으로 학습시킨 모델을 사용하여 semi-self supervised learning 방법을 사용하여 오토레이블링을 하는 방법 등 다양한 방법을 적용할 수 있다.

4.2 편견 발언 분류

<Table 3>에서 Bias 성능은 편견 분류 작업에 LR모델, KoBERT, KoELECTRA, KoGPT2, KcELECTRA을 적용한 성능이다. 각 모형은 73~84% 사이의 정확도를 보인다. 이는 혐오 표현 분류에서의 성능이 62~66% 정도였으므로 비교적 좋은 성능이라고 할 수 있다. <Table 1>을 보면 Bias classification dataset은 none 데이터가 gender bias, others bias 데이터 수의 합보다 3배 많은 매우 불균형한 데이터셋임에도 불구하고 혐오 표현 분류보다 높은 f1-score를 달성하였다. f1-score가 특히 낮은 others bias의 성능저하 요인을 파악하기 위해 KOCO dataset보다 bias label이 세분화된 Unsmile dataset을 추가로 사용하여 특히 분류가 어려운 bias label이 있는 것인지, gender 이외의 bias는 분류성능이 좋지 않은 것인지 등 bias label의 성능저하 요인을 분석한다. 본 절에서는 편견에 대한 분류 문제의 데이터 불균형이 심하므로 소수 클래스인 gender bias와 others bias의 성능 차이가 나는 이유를 분석하기 위해 각 레이블별로 성능에 영향을 끼친 요인을 분석하고, 편견 데이터셋의 데이터 레이블링 오류에 대해서도 살펴보고자 한다.

(1) 성별 관련 단어

클래스별 f1-score를 <Table 6>에서 확인하였을 때 gender bias는 71~75%의 비교적 좋은 성능을 보였다. 편견 분류 성능에 영향을 끼친 요인들을 살펴보기 위해서 LIME을 통해 모델

Tabel 4. Mislabeled hatespeech data

Comment	hate
10+8 진짜 이승기랑 비교된다	none
그래도 17 18 19회분은 보여주는게	offensive
설마 ㅈ 현정 작가 아니지??	hate

Input Text	True label	predict label
아직도 죽은 엄마 팔아먹어 관심 받으려고 하는 관중이네...	bias: none	bias: gender

Figure 4. Bias Detection

의 예측에 영향을 많이 준 단어를 분석하였다. 데이터 태깅 가이드라인에 따르면, 성별에 대한 편견적인 내용을 포함하는 문장이면 gender bias로 태깅하였다. 따라서 모델은 입력 데이터에서 여자, 남자 등 성별과 관련된 단어가 발견되면 gender bias로 분류하는 경향이 있었다. <Figure 4>는 KoBERT 모델의 bias detection 결과를 LIME으로 분석한 것이다. '엄마'라는 단어의 영향을 크게 받아 gender bias로 예측한다. gender bias는 성별과 관련된 단어를 보고 판단하면 되기 때문에 분류 성능이 좋지만, <Figure 4>처럼 특정 단어에 의존하는 경향이 있다. 따라서 특정 단어에 대한 의존성을 줄이기 위해 문맥 정보를 많이 반영할 수 있는 모형이 필요하고, 주어진 데이터셋의 데이터 불균형이 심하므로 none bias에 비해 약 5배 정도 적은 gender bias 데이터를 추가하여 충분한 학습 데이터를 제공할 필요가 있다. 또한, 성별과 관련된 단어가 들어가 있지만, 성별에 대한 편견적인 내용이 없는 데이터를 더 추가하는 등 데이터의 다양성 측면도 함께 고려할 필요가 있다.

(2) 잘못 레이블링 된 데이터

편견 발언 분류 모형의 예측 결과를 LIME으로 분석하고, 학습에 영향을 준 데이터들에 대해서 분석하였다. 본 연구에 사용된 데이터셋은 크라우드소싱 방법으로 레이블링되어 구축되었으므로 직접 데이터셋을 살펴보고 데이터 태깅 가이드라인과 다르게 태깅이 되어있는 데이터가 있는지 잘못 레이블링 된 데이터를 확인하였다.

<Table 5>의 첫 번째, 두 번째 예시 문장들은 외모에 대해 예쁘다는 수식어를 사용하여 none bias 데이터로 예측될 수 있지만, 이것은 성별에 대한 편견이 포함된 문장으로도 보일 수 있다. 세 번째, 네 번째 예시 문장은 '예쁘다', '천상여자'라는 발언이 gender bias가 있다고 labeling된 반면, 첫 번째 두 번째 문장에서는 비슷한 맥락으로 사용된 '예쁘다', '몸매 갑'이라는 발언이 gender bias가 없는 none bias 데이터로 labeling 되었다. bias 판단에 있어 더 명확한 기준을 제시하여 Mislabeled data를 줄인다면 모델이 더 명확하게 학습하여 더 좋은 결과를 기대할 수 있을 것이다.

(3) 다양한 편견 포함 발언

편견 발언의 클래스별 f1-score을 <Table 6>에서 살펴보았을 때 other bias 의 경우, 45~57% 정도의 비교적 낮은 성능을 보였다. 각 클래스별 분류 성능을 특정 bias가 아닌 여러 bias를 가지고 있는 others bias는 분류 성능이 좋지 않았다. <Table 3>의 모델별 bias classification에 대한 성능을 보면 모든 모델에서 others bias에 대한 분류 성능이 좋지 않을 것을 확인할 수 있다.

이러한 결과는 댓글 데이터에서 중복된 bias를 가진 데이터의 영향을 받은 것으로 판단된다. 입력 문장에서 인종, 장애 등 다양한 bias들이 중복되어 나타난 경우, label은 others bias이지만, gender bias와 다른 bias가 중복되어 나타난 경우, label이 gender인 경우가 많았다. <Table 7>의 예시처럼 label이 gender bias이지만, gender bias 외에 장애, 나이에 대한 bias가 포함되어있을 경우 others로도 분류될 수 있다.

편견 발언 분류 작업의 others bias의 성능이 비교적 낮은 이

Table 5. Mislabeled Bias Data

Comment	bias
아이구~ 지민씨 너무 예쁘다	none
손태영은 몸매 갑	none
송혜교보다 더 예쁘고 우아하다 오연수도 닮았고5kg만 더 빼면 최고 자연미인~♡♡♡♡	gender
두 분 권태기없이 이뿌게 사랑하며 아껴주는 모습 보기 좋아요~~사연언니 천상 여자여자~~~♡♡♡	gender

Table 6. Comparison of F1-score of Bias Classification

	LR	KoBERT	KoELECTRA	KoGPT2	KcELECTRA
gender	0.71	0.75	0.76	0.63	**0.77**
others	0.46	0.45	**0.57**	0.46	0.52

Table 7. Mislabeled Data as Gender Including Various Biases

Comment	bias
그냥 김치석 따로 만들어라. 지적 장애인 대우 해주께 ㅋㅋㅋㅋ	gender
애미나이 50세면 유산 및 기형아나 장애아 출산가능성 엄청높을듯	gender
40대 아주머니들 화가 많이 나셨네요	gender

Table 8. Comparison of F1-score of Unsmile Dataset using KoELECTRA and KcELECTRA

	KoELECTRA	KcELECTRA	number of data
여성/가족	0.72	0.80	394
남성	0.82	0.86	334
성소수자	0.84	0.87	280
인종/국적	0.77	0.84	426
연령	0.69	0.83	146
지역	0.83	0.91	260
종교	0.86	0.89	290
기타 혐오	0	0	134
악플/욕설	0.66	0.70	786
clean	0.73	0.77	935

유가 others bias에 다양한 bias가 포함되어 있어 분류가 어렵기 때문에 발생하는 것인지 확인하기 위해 bias label이 보다 세분화되어 있는 Unsmile dataset(Kang *et al.*, 2022)을 추가로 학습하였다. 편견 발언 분류에서 gender bias label의 f1-score가 가장 높은 KcELECTRA와 others bias label의 f1-score가 가장 높은 KoELECTRA 모델에 Unsmile dataset을 학습시켜 성능을 확인하였다. Unsmile dataset의 학습에서는 KcELECTRA 모델이 더 우수한 성능을 보였다.

Unsmile dataset은 특정 집단에 대한 편견을 세부적으로 labeling한 한국어 혐오표현 데이터셋이다. Unsmile dataset은 multi-label dataset으로 여러 bias가 포함된 문장의 경우 여러 개의 label을 갖는다. <Table 8>은 KoELECTRA와 KcELECTRA를 Unsmile dataset을 사용하여 multi-label classification을 학습시키고 성능을 평가한 결과이다. KoELETRA와 KcELECTRA에 KOCO dataset을 학습시켰을 때, others bias에 대한 f1-score는 각각 0.57, 0.52이다. Unsmile dataset 을 학습시킨 결과에서는 다양한 bias에 대해서 KOCO dataset 학습 시보다 더 높은 f1-score를 확인할 수 있다. 하지만, 기타 혐오 label에서는 매우 낮은 성능을 보인다. 이처럼 분류하고자 하는 bias가 명확하면 성능이 개선되고, '기타' 등으로 다양한 bias를 포함한 label일 경우 성능이 저하되는 것을 확인하였다.

따라서 한국어 혐오표현 데이터셋 구축 시에는 편견 및 혐오를 세분화하여 데이터셋을 구축하여야한다. KOCO dataset의 gender bias 데이터의 경우 gender bias와 others bias가 같이 나타나는 문장을 어떤 label로 분류할 것인지 명확한 가이드라인을 세워야 한다. 또한, KOCO dataset의 others bias는 보다 구체적인 label로 세분화할 필요가 있다. Unsmile dataset처럼 여성, 인종, 연령, 지역 등 label을 세분화하여 모델의 성능 개선을 기대할 수 있다.

5. 결 론

본 연구에서는 KOCO dataset을 LR모델, KoBERT, KoELECTRA, KoGPT2, KcELECTRA 모델에 학습시켜 모델별 분류 성능을 비교하고, 혐오 표현 탐지나 편견 탐지 모형에서의 예측 결과에 대해 정성적인 분석을 수행하여 악의적인 댓글 탐지 모형 개발을 위한 개선점을 제시하였다. 결과적으로 혐오 표현과 관련된 문장에서는 비속어나 신조어가 많이 쓰였지만, BERT 기반의 딥러닝 모형에서는 문맥 정보를 활용하여 예측을 수행하였기 때문에 OOV 문제의 영향이 크지 않다는 것을 발견하였다. 또한, 편견 탐지에 대한 분류 성능은 대체로 좋았지만, 성별과 관련 있는 단어에 지나치게 의존하여 '엄마'와 같이 성별이 포함된 단어가 있는 문장을 무조건적으로 gender bias로 분류하는 경우가 많았다. 또한 성별 이외의 기타 편견인 others bias 탐지에서는 매우 낮은 성능을 보였다. others bias가 연령, 종교 등에 대한 편견으로 KOCO dataset보다 더 세분화 된 Unsmile dataset에서 편견 분류에 더 좋은 성능을 보였으나 '기타 혐오' label에 대해서는 매우 낮은 성능을 보였다. 따라서 한국어 혐오표현 데이터셋 구축 시 편견 탐지 문제의 others bias와 gender bias의 경계를 명확히 하고, gender bias 외의 모든 bias를 아우르는 others bias 대신 성별, 연령, 종교 등 특정 bias를 명시하는 레이블링을 수행한다면 편견 탐지 성능이 향상될 것으로 기대된다. KOCO dataset은 크라우드 소싱을 통해서 레

이블링되었기 때문에 전체적으로 레이블링이 일관성 있게 되어있지 않아 모델의 예측에 영향을 준 경우가 많았기 때문에 라벨 노이즈를 고려한 방법론을 적용하는 등 해당 문제를 고려한 모형 구축이 필요하다. 본 연구에서 수행한 한국어 악성 댓글 탐지 모델 분석 결과는 향후 신규 데이터셋 구축이나 악성댓글 탐지 및 예방과 관련된 연구에 의미 있게 적용될 수 있을 것으로 기대된다.

참고문헌

Clark, K., Luong, M. T., Le, Q. V., and Manning, C. D. (2020), ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, In *International Conference on Learning Representations*.

Ha, Y. R., Cheon, J. S., Wang, I. S., Park, M. U., and Woo, G. (2021), A Filtering Method of Malicious Comments Through Morpheme Analysis, *Journal of the Korea Contents Association*, **21**(9), 750-761.

Hong, J. J., Kim, S. H., Park, J. W., and Choi, J. H. (2016), A Malicious Comments Detection Technique on the Internet using Sentiment Analysis and SVM, *Journal of the Korea Institute of Information and Communication Engineering*, **20**(2), 260-267.

Kang, T., Kwon, E., Lee, J., Nam, Y., Song, J., and Suh, J. (2022), Korean Online Hate Speech Dataset for Multilabel Classification: How Can Social Science Improve Dataset on Hate Speech?. arXiv e-prints, arXiv-2204.

Kim, J. S. (2018), Kor2vec, Github repository, https://github.com/naver/kor2vec.

Lee, J, B. (2021), KcELECTRA, Github repository. https://github.com/Beomi/KcELECTRA.

Moon, J. H., Cho, W. I., and Lee, J. B. (2022), BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection, *Proceedings of 8th international workshop on NLP for social media*, 25-31.

Park, J. W. (2021), KoELECTRA: Pretrained ELECTRA Model for Korean, Github repository. http://github.com/monologg/KoELECTRA.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016), Why should I trust you? Explaining the predictions of any classifier, In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Shin, M., Chin, H., Song, H., Choi, J., Lim, H., and Cha, M.(2021), “Hate Speech Detection in Chatbot Data Using KoELECTRA”, In *Annual Conference on Human and Language Technology*, 518-523.

SKT-AI KoGPT2 (2021), KoGPT2, Github repository, https://github.com/SKT-AI/KoGPT2.

SKTBrain KoBERT (2019), KoBERT, GitHub repository. https://github.com/SKTBrain/KoBERT.

저자소개

이세영 : 성신여자대학교 융합보안공학과에서 2021년 학사학위를 취득하고 성신여자대학교에서 미래융합기술공학과 석사과정에 재학 중이다. 연구분야는 기계학습, 자연어처리, 데이터마이닝이다.

박새롬 : 서울대학교 산업공학과에서 2013년 학사, 2018년 박사학위를 취득하였다. 2019년부터 성신여자대학교 융합보안공학과 조교수로 재직하고 있다. 연구분야는 인공지능 및 보안, 안정성 분석, Robust Training, 암호화를 통한 개인정보보호 머신러닝이다.