KcBERT 언어 모델을 사용한 뉴스 기사 혐오 댓글 분류

강지수 ○ 박하명 강승식
국민대학교 컴퓨터공학과
tn1078@kookmin.ac.kr, hmpark@kookmin.ac.kr, sskang@kookmin.ac.kr

Classification of Hate News Article Comments using KcBERT Language Model

Jisoo Kang○, Ha-Myung Park, Seungshik Kang
Department of Computer Science, Kookmin University

요 약

본 연구에서는 한국어 뉴스 댓글 기사를 활용하여 사전학습 임베딩 모델인 KcBERT(Korean comment BERT)를 사용하여 임베딩을 수행하고 뉴스 기사의 혐오 댓글을 분류하는 작업을 진행하였다. 실험 결과로 개별 문장, 즉 댓글 하나에 대한 예측을 보여주는 결과와 뉴스의 전체 댓글에 대해 혐오 판단을 하여, 해당 뉴스의 혐오 댓글 잠식률(전체 댓글에 대한 혐오 댓글 비율)을 확인하는 실험, 두 종류로 나누어진다. 두 결과 모두 높은 정확도를 보인다.

1. 서 론

최근 '혐오의 시대'라는 말이 생길 정도로 수많은 혐오 표현이 쏟아지고 있다. 특히, 인터넷 뉴스 기사의 경우 악플이 없는 기사를 찾아보기가 힘들 정도이다. 불특정다수를 향한 무차별적인 혐오 표현들은 보는 사람들로 하여금 불쾌감 뿐 아니라 피곤함까지 느끼게 한다[1,2]. 악플 문제를 해결하기 위해 몇몇 플랫폼에서는 특정 분야의 기사는 아예 댓글을 제한하는 등의 해결책을 제시하였으나, 모든 분야의 댓글 섹션을 폐지할 수 없고, 사람들의 의견을 자유롭게 나눌 수 있는 공간이 사라진다는 한계가 존재한다.

따라서 댓글 및 기사를 무조건적으로 차단하기보다는 사용자에게 선택권을 줌으로써 사용자가 보다 주체적으로 높은 질의 인터넷 기사 서비스를 즐길 수 있는 문제 해결 방안이 필요하다. 혐오 표현 탐지에 관한 많은 연구가 이루어져 왔으며, 탐지 모델을 주로 딥러닝 기법이 사용되고 있다[3-6].

본 연구에서 해결 방안으로 제시하는 모델은 혐오 표현을 판단하는 이진 분류 모델로, 기사의 댓글에 일정 퍼센트 이상 혐오 혹은 정치, 젠더 갈등과 같이 논쟁을 야기할 수 있는 표현이 다수 포함되어 있을 때, 혐오표현 탐지 정보를 제공하여 사용자가 해당 기사에 대한 혐오 관련성을 인지할 수 있도록 주의 표시를 띄우고자 하는 것이다. 연구를 진행하기 위해 데이터셋은 한국어 혐오성 표현 데이터셋[7]을 활용하였고, 언어 모델은 kcBERT 를 사용하였다[8].

2. 한국어 혐오성 표현 데이터셋

한국어 혐오 데이터셋은 뉴스 기사의 댓글이 혐오를 띄고 있는지, 띄고 있다면 어떤 종류의 혐오인지를 라벨링 한 데이터이다. 학습용 데이터셋은 train 데이터 7,896 개, dev 데이터 471 개, test 데이터 974 개로 구성되어 있으며, 테스트를 위해 라벨이 붙어있지 않은 예제 2,033,893 개를 Kaggle Competition 을 통해 테스트할 수 있다[9-11].

데이터 구조는 comments(뉴스 댓글), news_title(뉴스 제목), contain_gender_bias(성적 차별 여부), bias(차별 종류), hate(혐오 여부)로 구성되어 있다. contain_gender_bias 는 True, False 로 단순 여부만 판단하는 반면, bias 와 hate 는 카테고리형 데이터이기 때문에 모델 훈련 전 적절한 처리를 해주어야 한다. 혐오 데이터의 value 값은 표 1 에서 확인할 수 있다.

표 1. 혐오 라벨링 구조

Label	Value
contain_gender_bias	True, False
bias	none, others, gender
hate	none, offensive, hate

이때, hate 와 offensive 의 차이는 hate 의 경우 기존에 정의되었던 혐오 표현 및 모욕에 충분히 해당한다고 볼 수 있다 생각되는 표현들(특정 아이덴티티에 기반하여 개인, 그룹에 대한 적대감을 표출하는 혐오, 개인 및 그룹의

[1] 본 연구는 2022 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음 (2022-0-00964)

사회적 체면을 심각하게 손상할 수 있는 모욕)에 해당한다. 또한, 읽었을 때 기분을 상하게 하지만, 혐오 혹은 모욕 정도는 아니라고 생각되는 표현들에 대해 offensive 의 레이블이 책정되었다. 최종 레이블은 데이터 제작자들의 투표 및 회의로 진행되었으며 사람마다 인지하는 정도가 조금씩 다를 수 있다.

본 연구에서는 댓글의 혐오 여부만을 조사하므로, contain_gender_bias 라벨의 True, bias 라벨의 others, hate 라벨의 hate 만을 따로 '혐오 라벨링'으로 구분 짓고 나머지는 비혐오 라벨링'으로 분류하였다. bias 라벨의 others 만 혐오 라벨로 구분한 이유는, 모든 gender 는 이미 contain_gender_bias 이 True 이므로 이미 혐오 라벨링으로 구분이 되기 때문이다. 따라서, 전처리 과정에서 생략해 주었다.

3. KcBERT 를 이용한 혐오 표현 예측

3.1. KcBERT 언어 모델

기존 공개된 한국어 BERT 는 대부분 한국어 위키, 뉴스 기사, 책 등 잘 정제된 데이터를 기반으로 학습된 모델이기 때문에 실제로 NSMC 와 같은 댓글형 데이터셋, 정제되지 않았고, 구어체 특징에 신조어가 많으며, 오탈자 등 공식적인 글쓰기에서 나타나지 않는 표현들이 빈번하게 등장하는 데이터셋에 적용하기엔 어렵다는 문제를 가진다. KcBERT 는 이러한 데이터셋에 모델을 적용하기 위해, 온라인 뉴스에서 댓글과 대댓글을 수집하여 토크나이저와 BERT 모델을 처음부터 학습한 Pretrained BERT 모델이다. 모델의 학습 데이터는 2019.01.01~2020.06.15 사이에 작성된 댓글 많은 뉴스 기사의 댓글과 답글을 수집한 데이터이며, 텍스트만 추출 시 약 15.4GB, 1 억 1 천만개 이상의 문장으로 이루어져 있다.

본 연구의 목적인 '뉴스 기사 댓글 혐오 표현 분류'에 가장 적합하다고 판단하여 학습 모델로 선정하였다.

3.2. KcBERT 파인 튜닝

머신 러닝에서 파인 튜닝은 BERT, KcBERT 와 같이 사전 훈련된 모델을 새로운 데이터에 적합하도록 가중치를 조절하는 일종의 전이 학습에 대한 접근 방식이다. 본 연구에서는 네이버 뉴스 기사 댓글에 대해 사전학습(Pre-train)된 KcBERT 를 혐오 분류 모델에 적합하도록 한국어 혐오 표현 데이터를 사용하여 파인 튜닝 과정을 거친다. 이때, 사용한 데이터는 혐오 표현에 대한 라벨링을 완료한 전처리 된 데이터이다. KcBERT 의 경우, 사용자에게 편리하도록 NSMC Finetune Sample Code 가 Google Colab 으로 제공되고 있다. KcBERT 는 학습한 데이터셋의 크기에 따라 Base 와 Large 로 구분되는데 테스트를 위하여 Base 모델을 사용하였다.

3.3. KcBERT 를 사용한 혐오 표현 예측

혐오 표현 예측을 위해 KcBERT 에서 'predict_dataloader', 'predict_step' 함수를 구현하였다. 각각 예측 데이터 받아오기, 예측 진행을 위해 모델 실행하기 기능을 담당하는 함수이다. 두 함수를 구현함으로써 trainer.predict(model)이라는 모델에 대한 예측을 시행하는 코드가 동작한다. predict_dataloader 구현 시, 단일 예측, 즉 댓글 하나에 대한 예측을 진행하는 코드와 한 뉴스에 속하는 댓글 전체의 예측을 진행하는 코드를 따로 구현하였다. 이는 단일 예측 실험을 통해 적절한 혐오 표현 임계값을 구하고, 구한 값을 사용하여 뉴스 기사의 혐오 표현 잠식률(전체 댓글에 대한 혐오 댓글 비율)을 알아보기 위함이다. 이렇게 구한 혐오 표현 임계값은 0.7 이며 그림 1 에서 개별 예측 실험 결과를 근거로 설정한 값이다.

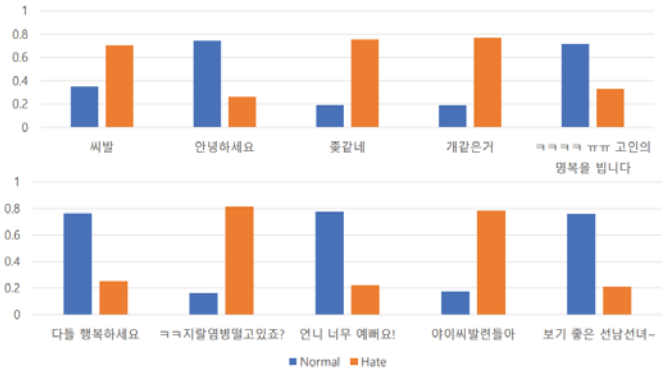

그림 1. 10 개의 문장에 대한 개별 예측 결과

그림 1 은 임의의 문장이 혐오성을 띄고 있는지에 대한 실험의 결과이다. 임의로 샘플링한 비속어가 섞인 문장 5 개("씨발", "좆같네", "개같은거", "ㅋㅋ지랄염병떨고있죠?", "아이씨발련들아")와 일상 문장 혹은 긍정적인 문장 5 개("안녕하세요", "ㅋㅋㅋㅋㅠㅠ고인의 명복을 빕니다", "다들 행복하세요", "언니 너무 예뻐요!", "보기 좋은 선남선녀~")에 대해 개별 예측을 진행한 결과, 10 개 문장 모두 모델이 제대로 예측하였음을 확인할 수 있다. 그림의 Y 축은 모델이 각 클래스(혐오, 정상)로 예측한 확률을 나타내며 주황색이 혐오 댓글, 파란색이 정상 댓글 클래스를 나타낸다.

표 2. 댓글 수가 가장 많은 10 개의 뉴스 샘플링

뉴스 기사 제목	댓글 수	잠식률
"'같이 살래요' 유동근, 장미희에 ""해아 물산 며느리, 내 딸이다"""	13	30%
"[종합]'뉴스룸' 김남주 ""타고난 연기자 아냐…악녀스런 고혜란 표현 고민"""	12	16%
"'아스달 연대기' 장동건-김옥빈, 들끓는 '욕망커플'→눈물범벅 '칼끝 대립'"	12	16%
"최종훈, 집단 성폭행 의혹…""동석했지만 성관계 NO"" [종합]"	12	75%
"'데뷔 12 년차 여유' 현아, 노출 사고에도 당당한 섹시 스타(종합)[Oh!쎈 이슈]"	12	56%
"[POP 이슈]박유천, 마약 혐의 시인→추가자백→""가족 보기 괴로워..풀려나고파""..."	11	72%
"'핸섬 타이거즈', 감격의 첫 승…손지창, 진심 조언 ""또 한 번의 기적 만들길..."	11	18%

"[SW 이슈] 월드와이드 방탄소년단, 대기록의 시작"	11	27%
"트와이스 미나, 韓 입국에 활동 복귀설+눈물..JYP 측 ""일정 참여 NO""[…	11	45%
"손현주, 이필모♥서수연 결혼식 사회 인증 "다시 뭉친 '솔약국집 아들들'""	11	30%

표 2 는 테스트셋에서 댓글 수가 가장 많은 뉴스 기사 10 개를 샘플링한 결과이다. 댓글이 가장 많은 뉴스는 13 개의 댓글을 갖고 있었으며 상위 10 개의 뉴스 모두 10 개 이상을 댓글을 갖고 있었고, 잠식률을 확인한 결과 약 0.9 의 정확도를 확인할 수 있었다.

4. 결론

한국어 혐오 데이터셋과 KcBERT 언어 모델을 활용하여 뉴스 기사에 달린 댓글의 혐오성 판단 분류 모델 생성을 목적으로 연구를 수행하였다. 진행한 예측 실험으로는 단일 예측, 즉 댓글 1 개에 대한 혐오 표현 예측을 시행한 실험과 이를 바탕으로 한 뉴스 기사에 속한 전체 댓글의 혐오 표현 예측, 즉 뉴스 기사의 혐오 댓글 잠식률을 확인하는 실험 총 2 가지를 진행하였다. 실험 결과, 두 실험 모두 높은 정확도를 보였다. 추가로 뉴스의 혐오 댓글 잠식률 실험에서 범죄 뉴스 기사에 대해 공통적으로 높은 잠식률을 보였는데, 이 점에서 착안하여 기사 제목의 혐오 예측 확률과 뉴스의 잠식률을 비교해 본 결과, 결과적으로 유의미한 연관 관계는 없었지만, 대체로 긍정적인 뉴스 기사 제목에서 혐오 표현이 아닌 정상 표현 예측률이 높음을 함께 확인하였다.

참고문헌

[1] H. Rizwan, M. Shakeel, and A. Karim, "Hate-speech and offensive language detection in Roman Urdu," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2512-2522, 2020.

[2] Z. Ahmed, B. Vidgen, and S. Hale, "Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning," EPJ Data Science, https://doi.org/10.1140/epjds/s13688-022-00319-9, 2022.

[3] P. Chiril, E. Pamungkas, F. Benamara, V. Moriceau, and V. Patti, "Emotionally informed hate speech detection: a multi-target perspective," Cognitive Computation, pp. 322–352, 2022.

[4] M. Karim, S. Dey, T Islam, S. Sarker, M Menon, K. Hossain, M. Hossain, and S. Decker, "DeepHateExplainer: Explainable hate speech detection in under-resourced Bengali language," in Proceddings of the IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1-10, 2021.

[5] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in Tweets," in Proceedings of the 26th International Conference on World Wide Web (WWW'17), pp. 759-760, 2017.

[6] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the Saudi Twittersphere," Applied Sciences, no.23 DOI:10.3390. app10248613, 2020.

[7] J. Moon, W. Cho, J. Lee. "Korean HateSpeech Dataset". Kocohub, https://github.com/kocohub/korean-hate-speech. 2020.

[8] KcBERT, https://github.com/Beomi/KcBERT.

[9] ko-nlp. "한국어 혐오 데이터셋 - Kopora". Kopora blog. https://ko-nlp.github.io/Korpora/ko-docs/corpuslist/korean_hate_speech.html

[10] J. Moon, W. Cho, and J. Lee, "BEEP! Korean corpus of online news comments for toxic speech detection," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 25-31, 2017.

[11] Korean Bias Detection, https://www.kaggle.com/c/korean-bias-detection.