KcELECTRA 를 활용한 혐오성 댓글 분류

이영준, 박하명, 강승식
국민대학교 컴퓨터공학과
younghe422@kookmin.ac.kr, hmpark@kookmin.ac.kr, sskang@kookmin.ac.kr

Classifying Hateful Comments with KcELECTRA

Youngjun Lee, Hamyoung Park, Seungshik Kang
[1]Department of Computer Science and Engineering, Kookmin University

요 약

본 논문에서는 KcELECTRA, KcBERT 등을 활용하여 혐오성 댓글 분류를 진행하였다. 사용 데이터셋은 한국어 혐오성 표현 데이터셋을 사용하였다. 한국어 댓글 데이터를 통하여 pretrain 한 KcBERT 와 KcELECTRA 를 둘 다 사용하여 비교했을 때, KcELECTRA 를 사용한 분류 모델이 좀 더 나은 성능을 보인 것을 알 수 있다. 이를 통해 특정 기사의 댓글의 몇 퍼센트가 악플에 잠식되었는지 파악할 수 있으며, 블라인드 기능을 추가시킬 수도 있다.

1. 서 론

현대 사회의 기사 댓글에서는 혐오 표현이나 댓글이 굉장히 많이 발생하는 것을 볼 수 있다[1,2]. 따라서 이러한 혐오 표현을 보고 싶어하지 않아하는 수요가 있으며, 이를 막기 위한 블라인드 기능이 필요하다[3-5]. 이 연구는 이러한 수요를 어느정도 해결해주고자 하는 목적을 가지고 있다. 이는 사용자가 특정 기사의 댓글창이 악플에 얼마나 잠식되었는지 파악하고자 하는 것이며, 이를 통해 사용자는 악플에 잠식된 기사를 블라인드하거나, 악플을 블라인드할 수 있다.

혐오성 분류 방법은 딥러닝 모델이 주로 사용되고 있다[6-8]. 본 연구에서는 특정 댓글이 혐오성 표현을 가지고 있는지 없는지 파악하는 이진 분류 문제를 다룬다. 이를 위해 pretrain 된 BERT 들을 이용하여 혐오성 표현 데이터셋에 맞게 finetuning 한 이후, 실제로 어느 정도의 결과가 나오는지 evaluation 을 진행하였다[9,10]. 데이터셋은 한국어 혐오성 표현 데이터셋을 사용하였으며, 이 데이터셋은 comment 데이터와 hate, bias, contain_gender_bias 의 라벨링 항목으로 구성되어 있다. 라벨링은 각각 혐오, 편견, 성 편견으로 나누어지며, 해당 프로젝트에서는 혐오 라벨링 태그를 사용하여 개발을 진행하였다.

2. 연구 방법

이를 위해 여러 언어 모델들을 사용하여, 그 중 가장 성능이 좋은 모델을 선택하여 최종적인 혐오성 댓글 분류 기능을 구현하였다. 분류 모델의 경우 여러 모델을 사용하였으나 그중 가장 결과가 좋은 KcELECTRA 를 사용하였다. 이 모델은 accuracy, precision, recall, F1 score 모두 90%가 넘는 결과를 보이고 있으며, 다른 모델들에 비해 혐오성 댓글을 탐지하는 성능이 우수함을 보이고 있다. 댓글 중 비교적 선플인지 악플인지 애매한 문장의 경우, 분류에 있어 틀리는 부분도 발생하며 이 부분에 대해서 고도화가 필요할 것으로 생각된다.

표 1. 여러 모델들간의 실험 비교

Model	Acc	F1	precision	Recall
KcBERT-base	0.844	0.894	1	0.808
KcBERT -large	0.813	0.870	1	0.769
KoBERT	0.688	0.783	0.9	0.692
distilKoBERT	0.563	0.696	0.8	0.615
KcELECTRA-base-v2022	0.906	0.943	0.926	0.962
KcELECTRA-small-v2022	0.719	0.816	0.870	0.769
KoBigBird-BERT-base	0.656	0.732	1	0.577

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

여러가지 임베딩 모델들의 평가 결과로부터 KcELECTRA 가 accuracy, F1 score, precision, recall 등 모든 평가척도에서 다른 모델들보다 나은 결과를 보이고 있음을 알 수 있다. 기존에 실험에서 진행하였던 KcBERT 의 경우, precision 은 1 임에 비해 Recall 은 0.808 로 큰 차이를 보이고 있으며, 이는 모델이 true 라고 예측한 것은 모두 true 가 맞지만, 실제로 true 인 것 중 모델이 판별해낸 것은 80% 정도임을 의미하며, false 로 판별한 것들 중에 True 가 포함되어 있음을 나타낸다.

반면에 KcELECTRA 는 precision 과 recall 모두 90% 이상의 결과를 보이며 이러한 문제를 해소한 것을 알 수 있으며, 다른 모델들에 비해서도 상당히 우수한 결과를 내고 있어 해당 모델을 통해 혐오성 댓글 분류를 진행하였다.

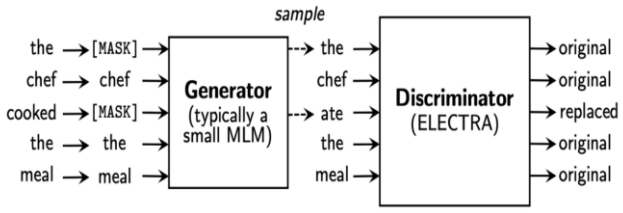

그림 1. ELECTRA 모델의 구조.

사용 모델로 지정한 KcELCTRA 모델은 한국어 데이터를 활용해 학습시킨 모델로, 다중언어로 학습된 모델보다 한국어 처리에 훨씬 뛰어난 성능을 보인다. 우선 KcBERT 와 같이 한국어 댓글 데이터를 사용해 학습하였으며, 그 때문에 댓글과 같이 잘 정제되어 있지 않고 비속어나 오탈자, 신조어가 많은 데이터셋을 학습하는데 적합하다고 볼 수 있다. 또한 BERT 와는 학습 방식이 다른데, pretrain 할 때 문장의 masking 된 부분을 맞추는 방식으로 학습하는 BERT 기반 모델들과 달리, ELECTRA 는 특정 토큰을 masking 하는 것이 아닌 generator 를 통해 그럴듯한 가짜 토큰으로 대체하고, 이를 discriminator 가 판별하는 방식으로 학습이 진행된다. BERT 의 경우 마스킹을 15%의 데이터만 진행하기 때문에 원하는 성능이 나오기 위해서는 많은 데이터가 필요한 반면, ELECTRA 의 경우 모든 토큰에 대해 진행되기 때문에 더 효율적인 방식이라고 할 수 있다.

3. 연구 결과

KcELECRA 를 문장에 적용하였을 때는 표 2 와 같은 결과를 보이고 있으며, 확실히 부정적인 요소나 혐오성 표현을 가진 악플이나, 긍정적인 요소를 가진 선플의 경우를 잘 분류해 내는 것을 볼 수 있다. 반면 마지막 문장처럼 선플인지 악플인지 애매한 문장의 경우 분류를 확실히 하기 쉽지 않은 면이 있으며, 따라서 negative 에 대한 threshold 를 0.675 로 선정하여 확실히 악플인 것들만 걸러내는 방식으로 연구를 진행하였다.

표 2. 혐오성 댓글 분류 예

Comments	Hate	Positive	Negative
1. 사람 얼굴 손톱으로 긁은것은 인격살해이고 2. 동영상이 몰카냐? 메걸리안들 생각이 없노	True	0.0481	0.9519
힘내소...연기로 답해요.나도 53살 인데 이런일 저런일 다 있더라구요.인격을 믿습니다..홧팅	False	0.6036	0.3964
6 명이서 그깟 20 년 한거가지고 참 ㅋㅋㅋㅋㅋㅋㅋㅋ강필주는 혼자 20 년 기다리고 참아왔다..	True	0.0795	0.9205
ㅉㅉ 아빠 없이 무슨.... 친가 뿌리와 족보는 제대로 알려주고 키울지 걱정이구먼	True	0.0404	0.9596
1,2 화 어설펐는데 3,4 화 지나서부터는 갈수록 너무 재밌던데	False	0.4860	0.5140

따라서 이러한 과정을 거쳐 만들어진 모델을 이용하여 특정 기사 댓글의 몇 퍼센트가 혐오성 댓글인지 알 수 있다. 또한 특정 기사 댓글의 몇 퍼센트가 악플로 잠식되어 있는지 구체적인 수치를 나타낼 수 있으며, 이를 이용해 기사 댓글을 보기 전 블라인드할 수 있다.

표 3. 뉴스기사 댓글에 적용 예

News	Output
"'같이 살래요' 유동근, 장미희에 ""해아 물산 며느리, 내 딸이다"""	69%
"최종훈, 집단 성폭행 의혹…""동석했지만 성관계 NO"" [종합]"	91%
"트와이스 미나, 韓 입국에 활동 복귀설+눈물..JYP 측 ""일정 참여 NO""[...	81%
"손현주, 이필모♥서수연 결혼식 사회 인증 “다시 뭉친 ‘솔약국집 아들들’""	18%

4. 결론

본 연구는 인터넷 댓글 데이터셋으로 pretrain 된 언어 모델을 한국어 혐오성 데이터셋으로 finetuning 하여 이진분류 task 를 수행하는 것이다. KcELECTRA 라는 모델을 이용하여 혐오성 댓글 분류를 진행하였으며, 기존에 사용하였던 KcBERT 보다 더 높은 성능을 낼 수 있었다. 또한 특정 기사의 악플이 얼마나 악플로 잠식되어 있는지 구체적인 수치를 제시할 수 있는 기능을 구현하였으며, 이를 통해 사용자들에게 혐오성 댓글을 블라인드 하고자 하는데 활용할 수 있다.

참고문헌

[1] J. Nobata, A. Tetreault, T. Mehdad, and Y. Chang, “Abusive language detection in online user content,” in Proceedings of the 25th International Conference on World Wide Web(WWW’16), pp. 145-153, 2016.

[2] 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.

[3] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, “Gender bias in neural natural language processing,” in Nigam V. et al. (eds.) Logic, Language, and Security, LNCS, vol. 12300, pp. 189-202, 2020.

[4] 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.

[5] 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.

[6] 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.

[7] N. Mullah and W. Zainon, “Advances in machine learning algorithms for hate speech detection in social media: a review,” IEEE Access 9(88):364–388. https://doi.org/10.1109/ACCESS.2021.3089515, 2021.

[8] P. Fortuna, J. Soler-Company, and L. Wanner, “How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?,” Information Processing & Management 58(3):102524 DOI 10.1016/j.ipm. 2021.102524, 2021.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of deep bidirectional transformers for language understanding,” CoRR abs/1810.04805, 2018.

[10] K. Clark, M. Luong, Q. Le, and C. Manning, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” International Conference on Learning Representation, Advance online publication. https://arxiv.org/abs/2003.10555, 2020.