KcBERT를 활용한 한국어 악플 탐지 분석 및 개선방안 연구

정세영°[1], 김병진°[1], 김대식°[1], 김우영[1], 김태용[1], 윤현수[2], 김우주*[1]
[1]연세대학교 스마트 시스템 연구실, [2]연세대학교 산업공학과
{tpdud3406, jin_kbj, kds8266, timothy, kasamdi5, hs.yoon, wkim}@yonsei.ac.kr
°공동 1저자, *교신저자

Analyzing Korean hate-speech detection using KcBERT

Seyoung Jeong°[1], Byeongjin Kim°[1], Daeshik Kim°[1], Wooyoung Kim[1], Taeyong Kim[1], Hyunsoo Yoon[2], Wooju Kim*[1]
[1]Smart System Lab, Yonsei University, [2]Dept. of Industrial Engineering, Yonsei University
°First Author, *Corresponding Author

요 약

악성댓글은 인터넷상에서 정서적, 심리적 피해를 주는 문제로 인식되어 왔다. 본 연구는 한국어 악성댓글 탐지 분석을 위해 KcBERT 및 다양한 모델을 활용하여 성능을 비교하였다. 또한, 공개된 한국어 악성댓글 데이터가 부족한 것을 해소하기 위해 기계 번역을 이용하고, 다국어 언어 모델(Multilingual Model) mBERT를 활용하였다. 다양한 실험을 통해 KcBERT를 미세 조정한 모델의 정확도 및 F1-score가 타 모델에 비해 의미 있는 결과임을 확인할 수 있었다.

주제어: 한국어 악성댓글 탐지, KcBERT, 기계번역, 다국어 언어 모델, mBERT

1. 서론

악성 댓글(이하 '악플')은 인터넷상에서 정서적, 심리적 피해를 주는 문제로 인식되고 있으며, 다양한 형태와 표현 방식을 가지고 있어 탐지 모델의 개발은 도전적인 과제이다. 악플 탐지 모델을 개발하기 위해 사전 학습(pretrained)된 언어 모델들을 활용할 수 있다. 대표적으로 Transformer [1] 아키텍처를 적용한 BERT [2]가 있다.

본 연구에서는 한국어 악플 탐지 모델의 베이스라인으로 KcBERT [3]를 활용하였다. 그리고 비교를 위해 여러 모델을 활용하여 KcBERT와 성능을 비교하였다. 그리고 한국어 악플 탐지 모델의 성능을 높이기 위해 데이터 증강 및 다국어 모델 활용한 실험을 추가로 진행했다.

2. 관련연구

인터넷 사용이 증가함에 따라 관리자가 악플 관리를 직접 하기 어렵기 때문에 악플 탐지 모델 연구가 이루어져 왔다. 다양한 자연어처리 연구에서 Transformer가 좋은 성능을 보임에 따라 한국어 악플 탐지에도 이용됐다. 일반적으로 사전 학습(pretrained)된 Transformer를 적용한 BERT, GPT 등을 활용하여 한국어 악플을 탐지하였다.

KoBERT [4]는 다국어 언어 모델(Multilingual Model)인 mBERT(multilingual BERT)에 한국어 위키피디아 500만 문장, 그리고 한국어 뉴스 데이터 2,000만 문장을 추가로 사전 학습한 모델이다. 따라서 기존 mBERT와 달리 한국어 위키를 기반으로 학습이 된 토크나이저(SentencePiece)를 사용하였다. 한국어의 언어적 특성을 학습하였기 때문에 한국어 악플 탐지 모델에 적용할 수 있다. 하지만 이러한 모델은 위키피디아 등의 문서 데이터를 기반으로 학습되어 있기 때문에 신조어 등의 비정형 데이터를 잘 파악하지 못할 수 있다. 따라서 한국어 댓글을 기반으로 학습이 된 모델을 활용하여 악성 댓글 탐지에 적용할 필요가 있다. 대표적으로 KcBERT [3]가 있다. KcBERT는 사람들이 작성한 댓글이나 신조어 등의 비정형 데이터에도 대응할 수 있도록 2019년 1월부터 2020년 6월까지의 네이버 뉴스 기사의 댓글 약 1억 1천만 건의 데이터로 학습된 모델이다.

다양한 언어를 활용하기 위해 다국어 언어 모델을 활용할 수 있다. 다국어 언어 모델은 다양한 언어와 대량의 데이터로 사전 학습된 모델로, 교차언어 전이 학습(cross lingual transfer learning)[5]을 이용하여 영어로 미세조정 하여도 한국어 악성댓글 탐지에 활용할 수 있는 장점이 있다. 대표적으로 mBERT [6]는 104개의 언어로 구성된 위키피디아 데이터를 BERT로 학습시킨 다국어 언어 모델이다. 즉, mBERT는 다양한 언어로 학습되었기 때문에 언어 간 특성을 잘 파악한다는 장점이 있다. 따라서 한국어 악플 탐지를 개선하기 위해 다국어 언어 모델인 mBERT를 하나의 후보 모델로 선정했다.

3. 실험 방법

3.1 데이터셋

본 연구에서는 학습 및 평가 데이터로 kocohub[1] 데이터셋과 Korean UnSmile[2] 데이터셋을 활용하였다.

kocohub 데이터셋은 총 9,381개의 라벨링 된 데이터로, 7,896개의 훈련 데이터셋과 471개의 검증 데이터셋, 그리고 974개의 테스트 데이터셋으로 구성되어 있다. 혐오 표현과 관련해서 hate, offensive 그리고 none 라벨을 제공한다. 본 연구에서는 hate와 offensive는 악플로, none은 악플이 아닌 데이터로

[1]https://github.com/kocohub/korean-hate-speech
[2]https://github.com/smilegate-ai/korean_unsmile_dataset

활용하였다.

Korean UnSmile 데이터셋은 총 18,742개로 혐오 표현은 10,139개, 악플/욕설은 3,929개, Clean은 4,674개로 구성되어 있으며, 혐오 표현은 제외하고 악플/욕설 데이터는 악플로, Clean 데이터는 악플이 아닌 데이터로 활용하였다. 따라서 두 데이터셋을 전처리한 후 총 16,745개로 재구성하였다.

3.2 모델

본 연구를 진행하기 위해 아래와 같이 모델을 구축하여 실험을 진행하였다. 실험으로 LSTM 기반 모델 2개, KoBERT 기반 모델 2개, KcBERT 기반 모델 2개를 활용하여 성능을 비교하였다. 실험으로 BiLSTM만 사용한 모델과 CNN, BiLSTM, LSTM을 결합한 구조를 활용하였다. 또한, BERT 기반 모델인 KoBERT를 미세 조정하거나, 미세 조정된 KoBERT에 BiLSTM을 결합하여 학습을 진행하였다. 마지막으로 KoBERT를 활용하였다. KcBERT 역시 미세 조정하거나, 미세 조정된 KcBERT에 BiLSTM을 결합하여 학습을 진행하였다.

- BiLSTM - MLP
- CNN - BiLSTM - LSTM
- KoBERT – MLP
- KoBERT - BILSTM
- KcBERT - MLP
- KcBERT – BiLSTM

3.3 실험 및 평가

3.3.1 실험 설정

데이터는 훈련(70%), 검증(10%), 평가(20%)로 나누어 사용하였다. 토크나이저의 경우 BiLSTM기반 모델과 KcBERT기반 모델은 WordPiece 토크나이저를 사용하고, KoBERT 기반 모델은 SentencePiece 토크나이저를 사용하였다. 손실함수는 이진 교차 엔트로피를 이용했으며 옵티마이저는 Adam을 이용하였다. 배치 사이즈는 32, 에포크는 20이며, max length는 60으로 설정하였다.

3.3.2 평가 및 분석

실험 결과를 확인하기 위해 평가지표로 정확도와 F1-Score를 사용하였다.

표 1은 KcBERT와 다른 모델을 비교한 결과이다. 이를 통해, KcBERT를 사용한 모델의 성능이 가장 높게 나온 것을 확인할 수 있다. 특히 KcBERT를 미세 조정한 모델의 성능이 가장 높았으며, 이는 한국어 댓글로 사전 학습되어 있기 때문에 한국어 댓글 특성을 잘 파악하여 악플 역시 잘 파악하였다고 판단된다. 따라서, KcBERT 모델을 활용하여 한국어 악플 탐지 모델을 개선해 보고자 한다.

표 1. 모델별 정확도 및 F1-Score

	정확도	F1-Score
BiLSTM - MLP	0.6838	0.6965
CNN - BiLSTM - LSTM	0.6530	0.6438
KoBERT - MLP	0.5984	0.5755
KoBERT - BiLSTM	0.5861	0.5838
KcBERT - MLP	0.7999	0.8061
KcBERT - BiLSTM	**0.8026**	**0.8098**

4. 개선 방안 연구

본 연구에서는 현재 공개된 한국어 악플 데이터가 부족한 것을 해소하고, 나아가 악플 탐지 성능을 높이기 위한 방안으로 데이터 증강에 대한 실험을 진행하였다. 첫 번째로, 데이터 증강을 위해 영어 악플 데이터를 기계 번역하였고, 두 번째로 다국어 언어 모델을 활용하였다.

4.1 번역 데이터 활용

본 연구를 위해 Hate Speech and Offensive Language [7] 데이터셋을 네이버 인공지능 번역기 '파파고'[3]를 활용하여 기계 번역하였다. 번역된 영어 데이터와 3.장에서 학습 데이터셋으로 사용했던 한국어 데이터에 추가하여 사용했다. 이 데이터를 학습에 활용하고, 데이터 증강 방법이 한국어 악플 탐지에 효과적인지를 확인하기 위해 평가의 경우 3.장에서 활용한 평가 데이터셋을 동일하게 사용하였다. 그리고 3.장에서 KcBERT가 악플 탐지에 있어서 타 모델보다 성능이 좋았으므로 KcBERT 기반의 모델을 비교 분석하였다.

4.1.1 실험 및 평가

실험을 위해 한국어는 3.장에서 활용한 데이터셋을 사용하였고, 한국어로 번역된 Hate Speech and Offensive Language는 전체 데이터셋 중 랜덤 샘플링하여 19,826개를 학습데이터로 사용하였다. 그리고 실험 시 max length, 토크나이저, 옵티마이저 등은 3.장 실험과 동일하게 사용하였다.

표 2. 데이터셋별 정확도 및 F1-Score

	정확도		F1-Score	
데이터셋	한국어	한국어+기계번역	한국어	한국어+기계번역
KcBERT	0.7999	**0.7883**	0.8061	**0.7948**
KcBERT - BiLSTM	**0.8026**	0.7874	**0.8098**	0.7927

표 2는 데이터 증강을 통해 한국어 악플 탐지 성능을 비교한 표이다. 그 결과 KcBERT를 미세 조정하거나, 미세 조정된 KcBERT와 BiLSTM을 결합한 모델의 정확도 및 F1-Score가

[3]https://papago.naver.com/

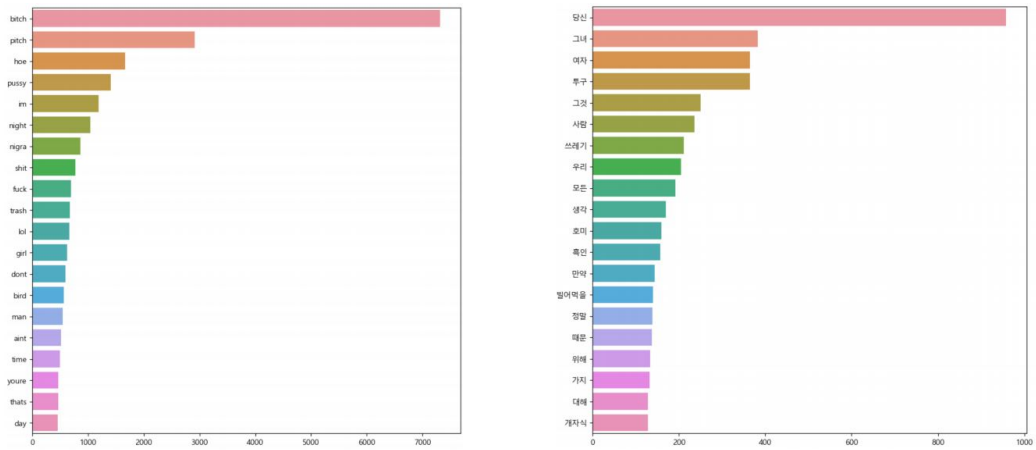

그림 1. 번역 전과 후 영어 데이터 명사 빈도수

비교적 높게 나왔다. 하지만 기계 번역을 통해 한국어 데이터를 증강해 성능 향상하고자 하였으나, 기대와 달리 성능이 더 떨어졌다. 이는 댓글이라는 특성상 신조어와 구어체, 비속어 등 정제되지 않은 구문들이 본래의 의미 및 작성자의 의도를 잃고 오역되었기 때문으로 생각된다. 이를 확인하기 위해 번역 전 영어 데이터 명사 빈도수를 확인해 보았다.

그림 1에서 번역 전 명사 빈도수를 통해 'bitch', 'pitch', 'hoe' 등의 순서로 빈도수가 높음을 알 수 있다. 'bitch'는 욕설이며 'pitch'와 'hoe'의 경우 은어 및 욕설과 유사한 발음의 단어로 욕설의 의미를 내포하는 단어이다. 하지만 번역 후의 명사 빈도수를 확인해 보니, '당신', '그녀', '여자' 등의 순서임을 알 수 있었다. 'pitch'와 'hoe'의 경우 '투구', '괭이'로 해석되었다. 이처럼 욕설, 은어, 욕설과 유사한 발음의 단어 등이 번역을 거치면서 본래의 뜻과 의미와는 다르게 번역되는 것을 알 수 있었다. 예시로, "bad bitches is the only thing that i like"는 "나는 멋진 여자들만 좋아해"라는 뜻인데 "나쁜 투구는 내가 유일하게 좋아하는 것이야"라고 해석되었고, "Yeah I'm waiting on that hoe Mayne!"는 "응 나는 그 여자를 기다리고 있어 메인!"이라는 뜻인데 "이야. 나는 그 괭이 위에서 기다리고 있어 아마!"라고 해석되었다. 이처럼 많은 욕설이나 신조어, 은어 등이 제대로 된 번역이 이루어지지 않았고, 이러한 원인으로 영어 데이터를 번역하여 학습 데이터의 수를 증강한 것이 모델 성능 향상에 도움이 되지 못했던 것으로 생각된다. 추가로 KcBERT에 BiLSTM 및 1D-CNN을 결합하는 경우 성능 개선이 있을 것이라 기대하였지만 미비하였다. 이는 KcBERT 자체의 성능이 좋아서도 있겠지만 Task 자체가 악플임을 구분하는 단순 이진 분류였기 때문으로 판단된다.

4.2 다국어 언어 모델 활용

다국어 언어 모델은 한국어뿐만 아니라 영어 역시도 학습 데이터로 활용할 수 있기 때문에 한국어 데이터 부족에 대한 문제점을 해소할 수 있으리라 생각하였다. 데이터는 4.장에서 활용한 Hate Speech and Offensive Language [7] 데이터로 학습을 진행한 후, 평가의 경우 3.장에서 활용한 평가 데이터셋을 사용하였다. 그리고 다국어 언어 모델로 교차언어 전이학습 효과가 있다고 보고된 mBERT를 사용하였다.

4.2.1 실험 및 평가

실험 설정은 실험 시 max length 설정, 토크나이저, 옵티마이저 등은 3.,4.장과 동일하게 하였다.

표 3. mBERT 정확도 및 F1 Score 산출 결과

	정확도		F1-Score	
데이터셋	영어	한국어+영어	영어	한국어+영어
mBERT	0.5330	0.7834	0.4730	0.7431

표 3을 통해 mBERT의 정확도 및 F1 Score를 확인할 수 있다. 표 1, 2의 결과보다 성능이 좋지 않음을 확인하였다. 즉, 영어 데이터를 확보해서 다국어 언어 모델을 활용하는 것이 성능 향상에 도움이 되지 않았음을 알 수 있다. 또한, mBERT를 한국어 데이터만을 활용하여 추가로 학습한 모델인 KoBERT보다도 성능이 나오지 못한 것을 확인하였다. 이는 영어 악플과 한국어 악플의 구조적 유사성과 언어적 특성이 달라 한국어 악플 특성이 반영되지 못했다고 판단된다. 따라서 다국어 언어 모델을 활용하는 것보다는, 도메인에 적합한 사전 학습 모델 및 데이터로 학습하는 것이 더 유의미하다고 생각된다.

4.3 결론

한국어 악플 탐지 모델 개발을 위해 모델별 정확도와 F1-Score를 산출한 결과 KcBERT 기반 모델의 성능이 가장 좋았다. 따라서 KcBERT 모델을 활용하여 한국어 악플 탐지 모델을 개선해 보았다. 먼저, 영어 데이터를 기계 번역하여 데이터를

증강하여 실험을 진행하였다. 하지만, 언어의 구조적, 문화적 차이로 인해 번역이 제대로 이루어지지 않아 성능이 향상되지는 않았다. 즉, 유의미하지 않은 데이터로 증강했기 때문에 성능 향상에 도움이 되지 않았다고 판단된다. 그리고 다국어 언어 모델을 활용하여 성능 향상을 기대했으나, 이 또한 성능 향상으로 이끌지 못했다. 즉, 영어 악플 데이터로 미세조정이 된 다국어 언어 모델이 한국어 악플의 언어적, 구조적 특성을 잘 반영하지 못했다고 생각된다. 하지만 뉴스 기사와 같은 정제된 말뭉치에 대해서는 번역을 통해 데이터를 늘리는 방안은 성능 향상에 도움이 될 것으로 기대된다.

감사의 글

이 논문은 국토교통부의 스마트시티 혁신인재육성사업으로 지원되었습니다.

본 연구는 연세대학교 딥러닝이론및응용(IIE7721)의 연장으로 진행되었으며, 윤현수 교수님께 감사드립니다.

참고문헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, Vol. 30, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] J. Lee, "Kcbert: Korean comments bert," *Annual Conference on Human and Language Technology*, pp. 437–440, 2020.

[4] S. TBrain, "Korean bert pre-trained cased (kobert)," *https://github. com/SKTBrain/KoBERT*, 2019.

[5] W. Kim, C. Jo, M. Kim, and W. Kim, "Marvelous agglutinative language effect on cross lingual transfer learning," *arXiv preprint arXiv:2204.03831*, 2022.

[6] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[7] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of the international AAAI conference on web and social media*, Vol. 11, No. 1, pp. 512–515, 2017.