KoBERT 모델 기반 한국어 뉴스 기사 제목 선정성 및 폭력성 검출

김민지[1], 김환도[2], 봉지민[3], 김대환[1]
[1,2,3]울산대학교 IT융합학과 학부생
[1]울산대학교 IT융합학과 조교수

spot4321@mail.ulsan.ac.kr[1], uiop889@ mail.ulsan.ac.kr[2], woochichi15@ mail.ulsan.ac.kr[3],
daehwankim@ulsan.ac.kr[1]

Detection of sexuality and violence in Korean news article title based on KoBERT mode

Min-Ji Kim[1], Hwan-Do Kim[2], Ji-Min Bong[3], Dae-Hwan Kim[1]
[1,2,3,1]Dept. School of IT Convergence, University of Ulsan

요 약

최근 선정적이고 폭력적인 뉴스 기사 제목의 여과 없는 노출로 인하여 유해한 언어 접촉이 빈번히 이루어지고 있다. 자극적인 단어에 지속적으로 노출되는 것은 인지 능력에 부정적 영향을 주는 것으로 알려져 있다. 따라서 이를 사전에 판별하여 정보를 수용하는 것이 필요하다. 본 논문에서는 KoBERT를 기반으로 한국어 뉴스 기사 제목에서 선정성과 폭력성을 검출하고자 한다. 학습을 위한 뉴스 기사 제목들은 인터넷에서 무작위로 총 9,500개의 데이터를 크롤링 하여 수집하였고, 모델의 말단에 NLNet을 추가하여 문장 전체의 관계를 학습했다. 그 결과 선정성 및 폭력성을 약 89%의 정확도로 검출하였다.

1. 서론

최근 스마트폰 및 인터넷이 보급됨에 따라 뉴스 기사를 누구나 손쉽게 접할 수 있게 되었다. 하지만 단편적인 제목으로 눈길을 끌어야 하는 뉴스 기사의 특성상 선정적이고 폭력적인 제목들이 사용자에게 여과 없이 노출된다. 이러한 유해한 언어들에 지속적으로 노출되는 것은 감각과 지각 능력에 부정적인 영향을 끼치는 것으로 알려져 있다. 따라서 선수적으로 기사에 포함된 내용을 판단하여 분별력 있게 수용하는 것이 필요하다.

본 논문에서는 인터넷에서 수집한 한국 뉴스 기사 제목 데이터들과 KoBERT 모델[1]을 활용하여 뉴스 기사 제목에 포함된 선정성/폭력성을 검출하고자 한다. 여기서 해당 모델에 NLNet (Non-Local Neural Networks) [2]를 추가하여 문장 내의 단어 관계를 반영해 정확도 향상을 도모하였다.

2. 관련 연구

2017년 구글에서 Transformer[3] 구조를 발표한 것을 기점으로, 사전 학습과 미세 조정을 통해 학습하는 방식의 언어 모델이 등장하였다. 특히 Transformer 기반의 언어 모델인 GPT-3 (Generative Pre-trained Transformer)[4], BERT (Bidirectional Encoder Representations form Transformer)[5]가 자연어 처리 분야에서 두각을 보인다. BERT는 방대한 양의 데이터인 Book Corpus(8억 개)와 Wikipedia(25억 개) 등을 이용하여 사전 학습이 진행되어 언어에 대한 이해도가 높고 여러 층의 Transformer 인코더를 쌓아 올린 양방향 모델이기 때문에 입력된 문맥 특성을 활용한다.

하지만 BERT와 같이 다른 언어로 학습된 언어 모델에 한국어 데이터 셋을 적용하는 것은 성능에 한계가 있다. 따라서 KoBERT 등 한국어 데이터 셋을 사전 학습 시킨 언어 모델을 만들기 위한 연구가 활발하게 진행 중이다.

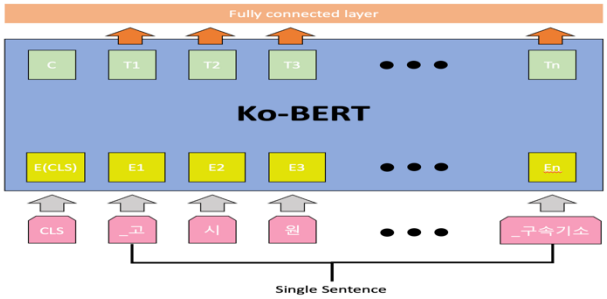
3. 모델 구조

본 연구는 한국어 뉴스 제목 분석을 목표로 하여 한국어 텍스트에 대한 성능을 높이기 위해 KoBERT 모델을 사용하였다. KoBERT 모델은 SKTBrain에서 공개한 모델로 한국어 위키 5백만 문장 및 단어를 학습하였다. 기본적인 구조는 BERT 모델과 유사하지만 한국어의 불규칙한 변화 특성을 반영하기 위한 SentencePiece 기반의 Tokenizer를 사용한다.

본 연구에서는 이에 더해 '괴물 공격수, 무차별적인 골로 대답한다' 등 폭력성을 띠는 단어를

포함하지만 내용은 폭력성을 포함하지 않는 문장들에 대한 정확도를 향상시키기 위해 모델의 뒷부분에 NLNet를 추가하였다. NLNet은 주로 이미지, 음성 분야에 활용되며 입력 시퀀스의 전체 구조를 고려하여 각 위치의 출력을 계산한다. 하지만 입력 시퀀스 내에서 전역 의미론적 관계를 학습하는데 효과적이기 때문에 자연어 처리 분야에서도 활용된다. 따라서 KoBERT 모델의 마지막 Self-Attention Layer를 NLNet으로 대체하여 문장 내에서 단어 관계를 파악하도록 하였다.

이외에도 epoch은 5, 글자의 최대 길이는 64, learning rate는 5e-5, optimizer는 AdamW로 설정하였다. 또한 모델 사용 목적이 분류이기 때문에 정확도 향상을 위해 출력층 앞에 fully-connected layer를 추가하였으며 모델의 일반화 능력을 향상시키기 위해 dropout 비율을 0.3으로 적용하였다.



<그림 1> KoBERT를 사용한 모델의 구조

4. 실험

4.1 데이터 수집 및 가공

본 연구에서 필요한 데이터가 기존에 존재하지 않아 웹 사이트에서 인터넷 뉴스 기사 헤드라인을 수집하여 사용하였다. 해당 데이터는 인덱스, 헤드라인 텍스트와 레이블로 이루어져 있으며 선정성(-1), 폭력성(1), 그리고 둘 다 관련되지 않음(0)을 기준으로 헤드라인을 키워드별로 수집하였다. 선정성 관련 키워드는 (선정성, 성매매, 성행위, 성희롱, 아동 성 착취 물, 음란물, 외설). 폭력성 키워드는 (폭행, 가해자, 공격 살인, 습격, 폭력, 학살). 일반 키워드는 (경제, 과학, 교육, 스포츠, 문학, 의료) 등을 선정하였다. 주제별로 약 3150개씩 총 9500개의 데이터를 수집하여 전처리를 진행했다. 학습 데이터와 시험 데이터의 비율을 8:2로 설정해 학습을 진행하였다.

<표 1> 수집 데이터셋 예시

본문	레이블
경찰, 행정복지센터서 난동 · 폭행 60대 구속 송치	1
"영화 '기생충' 가족같다" 20대 홍콩 모델 토막 살인 전말	1
학원생 상대 유사 성행위 학원장 징역 7년 선고	-1
"국내 남성 유튜버, 태국 현지 여성들과 선정적 방송"…경찰 조사 나서	-1
진천군, 봄 맞이 다양한 문화 행사	0
'영해 3 · 18독립만세 문화제'…영남 최대 독립만세운동 기리다	0

4.2 실험 결과

기존 KoBERT 모델과 NLNet를 추가한 모델로 총 2번의 실험을 진행하였다. 두 번의 실험 모두 직접 제작한 데이터 셋을 학습시켰다. 실험을 진행하여 도출된 정확도는 [표 2]와 같다.

<표 2> 모델의 정확도 비교

Model	Train set(%)	Test set(%)
KoBERT	89.76	87.16
KoBERT + NLNet	91.86	88.67

[표 2]를 보면 KoBERT 모델에 NLNet를 결합한 모델의 성능이 Train set 91.86%, Test sest 88.67%로 전반적으로 기존 모델보다 높은 정확도를 보이는 것을 알 수 있다. Train set과 Test set의 정확도 차이의 경우 기존의 KoBERT는 2.60%으로 NLNet을 추가한 모델의 3.19% 보다 나은 성능을 보였으나 미미한 수준이고 각 모델에서 검출 정확도의 차이 또한 크지 않아 각 모델의 한국어 뉴스 기사 제목의 선정성, 폭력성 검출이 성공적으로 진행되었음을 확인할 수 있다.

5. 결론

본 연구에서는 인터넷 기사의 제목들을 수집 및 평가하여 말뭉치를 제작하고 KoBERT 모델을 활용하여 뉴스 기사 제목에 있는 선정성과 폭력성을 검출하였다. 분류의 정확도를 높이기 위해 KoBERT 모델에 NLNet Layer를 추가하여 연구를 진행하였고 그 결과 훈련 데이터에서는 91.86% 테스트 데이터에서는 88.67%의 정확도를 도출해 모델의 분류 성능을 높였으며 폭력성, 선정성 검출에 대한 학습이 이루어졌음을 확인하였다.

실제로 수집한 데이터 중 판별에 실패한 데이터들을 분석해 보니 '미세먼지의 습격'이나 '살인 미소'등 문화적 은유에 따라 판단이 어려운 점이 있었다. 향후 연구에서는 모델을 개선하고 더 많은 데이터를 수집하여 한국어 특성에 맞도록 모델을 구축할 예정이다.

참고문헌

[1] SKTBrain, "KoBERT," GitHub repository, https://github.com/SKTBrain/KoBERT

[2] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He, "Non-local Neural Networks" CVPR, 2018

[3] Brown, T.B. et al., "Language Models are Few- Shot Learners," Advances in Neural Information Processing Systems, 33, 1877-1901, 2020.

[4]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, NIPS, Advances in Neural Information Processing Systems, 6000–6010,2017.

[5] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv, 2019.