딥러닝 기반 인공지능 생성 뉴스 탐지

장예훈[1]
[1]서강대학교 데이터사이언스·인공지능 학과 석사과정생
yhyh4420@sogang.ac.kr

A Study on Deep Learning-Based Detection of AI-Generated News

Ye-Hun Chang[1]
[1]Dept. of Data Science·Artificial Intelligence, Sogang University

요 약

생성형 인공지능의 발전으로 AI기자가 작성한 기사가 점차 증가될 것으로 전망되고 있다. 시간 절약, 경제성 등의 장점에도 불구하고 인공지능이 작성한 뉴스 내 허위정보 등으로 혼란이 사회적 문제로 제기되고, 이를 악용한 가짜뉴스 생성의 우려에 따라 구축모델의 필요성이 제기되고 있다. 이에 따라 실제 기사와 AI 작성 기사를 KoBART, KoELECTRA 모델과 두 모델을 앙상블한 모델에 적용시켰고, 그 결과 KoBART 모델의 Accuracy가 0.9995로 가장 높은 지표를 보였다.

1. 서론

최근 인공지능의 발전으로 AI기자가 작성한 기사가 점차 증가될 것으로 전망되고 있다.[1] 인공지능을 활용한 뉴스 작성은 신속한 콘텐츠 제작, 비용 절감 등의 장점이 있지만 잘못된 정보나 허위 사실이 포함된 뉴스를 작성하여 뉴스 소비자로 하여금 혼란을 가중시킬 수 있다. 특히 잘못 생성된 AI기사는 사회적, 정치적으로도 큰 파장을 일으킬 수 있다. 실례로 도널드 트럼프 공화당 대선 후보 피격 사건 이후 AI기자가 잘못된 기사를 생성해[2] 논란이 된 바 있다.

가짜뉴스에 대한 분류 연구는 활발히 진행되고 있다. Glove모델을 활용하여 99%이상 분류한 사례[3] 및 CNN, Fasttext 모델 활용 분류모델을 구현한 연구[4]를 확인할 수 있었고, 생성형 텍스트를 분류하는 연구도[5] 진행된 바 있다. 상기 연구들을 통해 한국어 텍스트에 대한 분류가 다양한 방식으로 진행됨을 확인할 수 있었다. 하지만 AI 생성 뉴스와 실제 뉴스를 분류하는 선행연구는 확인하기 힘들었다.

본 연구는 OPENAI API를 활용해 실제 뉴스 기반으로 뉴스를 재작성하여 실제 내용과 유사한 AI 뉴스 데이터를 확보하고, KoBART, KoELECTRA모델과 두 모델을 앙상블한 모델의 분류모델 성능 지표를 제시하고 결과를 비교 및 분석하고자 한다.

2. 활용 모델

본 연구에서는 KoBART과 KoELECTRA을 사용하였고, 앙상블은 soft voting방식을 적용하였다.

2.1. BART모델(Bidirectional and Auto-Regressive Transformer)

BART는 facebook AI연구팀에서 2019년에 발표한 Transformer기반의 시퀀스-투-시퀀스 모델이다. Pre-Training 과정에서 문장의 임의 토큰을 마스킹하는 방식으로 학습시키며, 이를 복원하는 학습과정을 통해 문맥을 이해할 수 있게 된다. 본 연구에서 사용하는 KoBART모델[6]은 SKT-AI에서 개발한 BART기반 모델로 모두의 말뭉치 v1.0, 청와대 국민청원 등 다량의 한국어 말뭉치로 학습하여 한국어 문장에 특화된 모델이다.

2.2 ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacement Accurately)

ELECTRA는 Google에서 공개한 BERT 후속모델로, 학습 효율을 높이기 위해 RTD(Replaced Token Detection)을 적용해 컴퓨팅 자원을 적게 소모하며 학습한다. RTD방식이란 학습 과정에서 일부 토큰을 fake token으로 바꾸고 Discriminator 과정을 거쳐 이를 탐지하는 이진 분류 문제를 해결하는 방식이다. KoELECTRA[7]는 한국어 뉴스, 위키, 나무위키, 모두의 말뭉치 등을 학습하여 한국어 문장에 특화된 ELECTRA 모델이다.

3. 실험 설계

3.1. 데이터셋

기존 뉴스 데이터는 AIHUB에서 제공하는 '뉴스 기사 기계독해 데이터'를 활용했다. 2021년 작성된 뉴스 제목과 뉴스 내용 400,056건의 데이터 중 학습 및 검증용 데이터로 20,000개의 뉴스 데이터를 사용했고, 최종 테스트 데이터는 8,000개의 뉴스 데이터를 사용했다.

AI 뉴스를 생성하기 위해 OPENAI API를 사용하였고, 생성 모델은 GPT-3.5-turbo를 사용하였다. GPT에게 프롬프트를 통해 기존 뉴스 제목과 내용을 재구성하라는 지시를 했고, 전처리를 통해 제목이나 내용을 생성하지 않은 데이터를 삭제하여 학습용 데이터 19,993개, 테스트용 데이터 7,949개를 확보하였다. 기존 사람이 직접 작성한 데이터와 생성한 데이터를 합쳐 학습 및 검증데이터 총 39,993개, 테스트용 데이터 15,949개의 데이터를 확보하였다.

AI뉴스데이터 생성 결과 기존 데이터와 <표 1>과 같은 차이를 보였다.

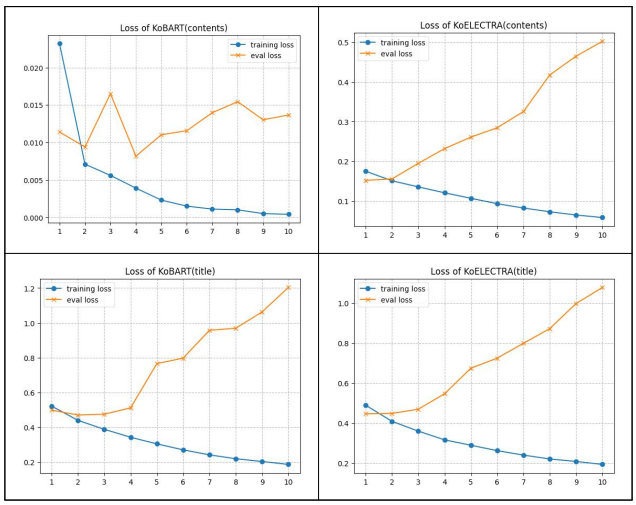
<표 1> 기존 뉴스 대비 AI 생성 뉴스 오류 비교

구분	기존 뉴스	AI 생성 뉴스
제목 오류	두산중공업, **연말까지 일부 휴업**… 400명 규모	두산중공업, 400명 **휴업으로 연말까지**
정보 삭제	인천시립박물관과 4개 분관**(송암미술관·검단선사박물관·한국이민사박물관·인천도시역사관)**도 재개관했다.	또한 인천시립박물관과 분관들도 재개관하여... (중략)
비정상적 종결	....알려져 뜨거운 반응이 **예상된다.**	피아노 연주로 뜨거운 반응 **예상.**
맥락 오류	영화 '그들만의 리그(A League of Their Own)'의 **실제 주인공이자 미국 여자 프로야구의 전설인 매리 플랫이 사망**했다. 향년 101세.	미국 여자 프로야구 전설 매리 프랫이 영화 '그들만의 리그(A League of Their Own)'의 **실제 주인공으로 알려졌으며**, 101세의 나이로 사망했다.
정보 오류	**코로나19 감독 대상 환자**(PDP) 가운데...	**코로나19 의심환자**(PDP) 중 ....

3.2. 실험 환경

Google Colab pro를 통해 모델 학습을 구현했으며, 해당 과정에서 GPU A100을 사용했다. 모델은 KoBART, KoELECTRA 모델을 사용했으며, 타이틀, 컨텐츠에 각각 학습시켰다. 해당 모델들을 10 epochs까지 학습데이터와 검증데이터를 4:1로 나누어 학습시킨 결과 대체적으로 epoch가 2일 때 검증데이터의 loss가 작은 경향을 보여 epoch는 공통적으로 2로 설정했다.

<그림 1> 제목, 내용별 각 모델의 train, eval loss 비교



통상 기사의 제목은 짧은 문장으로 이루어지고, 내용은 여러 문장으로 이루어지기 때문에 제목과 내용에 대한 분류성능을 각각 비교하기 위해 별개의 데이터로 학습시켰다. 학습 후 KoELECTRA모델과 KoBART모델의 softmax 결과를 soft voting하여 만든 앙상블 모델을 포함하여 기사 제목, 내용별 총 3개의 모델을 만들어 Accuracy, AUROC, Precision, Recall, F1 Score를 측정했다.

4. 실험 결과

실험 결과 각 모델별 Accuracy, AUROC, Precision, Recall, F1 Score를 도출하여 이를 비교할 수 있었다.

뉴스 내용에 대한 분류 결과는 다음과 같다.

<표 2> 각 모델별 생성형 AI 뉴스 분류 결과(내용 기준)

	KoELECTRA	**KoBART**	Ensemble
Accuracy	0.8029	**0.9995**	0.9992
AUROC	0.8845	**0.9999**	0.9996
Precision	0.7649	**0.9992**	0.9990
Recall	0.8729	**0.9997**	0.9995
F1 Score	0.8154	**0.9995**	0.9992

세 모델 모두 모든 지표에서 0.7 이상의 점수를 기록했다. 각 지표별 모델의 특징은 다음과 같다.

1) Accuracy : 세 모델 모두 0.8 이상의 정확도를 기록하여 생성형 AI 뉴스와 기존 뉴스를 분류함에 있어 우수한 성능을 보임을 확인했다. 특히 KoBART모델은 0.9995의 Accuracy를 기록하여 대부분의 생성형 AI 뉴스를 탐지함을 보였다.

2) AUROC : KoBART 모델의 AUROC가 0.9999로 가장 높았으며, 이는 본 모델의 전반적인 이진분류성능이 뛰어남을 보인다.

3) Precision : 모델이 AI 작성 뉴스라고 예측한 것 중 실제로 AI 작성 뉴스인 비율을 나타내는 지표로, 0.7649의 상대적으로 낮은 점수를 낸 KoELECTRA 모델과 비교하였을 때 KoBART모델은 0.9992의 점수를 보이며 이는 오분류를 거의 하지 않음을 시사한다.

4) Recall : 실제로 AI가 작성한 뉴스 중 모델이 AI 작성 뉴스라고 예측한 비율로, Precision과 마찬가지로 KoBART모델이 0.9997의 높은 점수를 기록했다. 이는 KoBART모델이 AI 작성 뉴스를 대다수 찾아냄을 의미한다.

5) F1 Score : F1 Score는 Precision과 Recall의 조화평균으로 KoBART모델이 가장 높은 점수인 0.9995를 기록했다.

뉴스 제목에 대한 분류 결과는 다음과 같다.

<표 3> 각 모델별 생성형 AI 뉴스 분류 결과(제목 기준)

	KoELECTRA	KoBART	Ensemble
Accuracy	0.4984	0.5015	0.5773
AUROC	0.5599	0.6290	0.6173
Precision	0.4984	0.0000	0.6033
Recall	1.0000	0.0000	0.4437
F1 Score	0.6652	0.0000	0.5113

뉴스 내용에 대한 분류성능보다 현저히 떨어지는 지표를 확인할 수 있었다. 각 모델의 Recall, F1 Score가 비정상적인 양상을 보여 기사 제목에 대한 KoELECTRA모델과 KoBART모델의 분류 능력이 없음을 확인할 수 있었다.

5. 결론

실험 결과를 통해 유의미한 AI 생성 뉴스기사 분류 모델을 구현할 수 있음을 확인했다. 특히 KoBART모델을 통해 구현한 모델의 경우 뉴스기사 내용을 높은 성능으로 분류할 수 있음을 각종 지표를 통해 확인하였다. 본 연구를 통해 향후 AI 생성 뉴스를 탐지함에 있어 유의미한 모델을 제시할 수 있음을 기대한다.

다만 제목에 대한 분류능력은 실험에 사용한 모든 모델이 현저하게 떨어지는 성능을 보여주었는데, KoELECTRA의 Recall, KoBART의 Precision, Recall, F1 Score과 같이 비정상적인 지표가 산출된 것을 보아 제목과 같은 짧은 문장의 분류를 학습하는데 있어 추가적인 연구 방법론이 필요할 것으로 예상된다.

본 연구에서는 GPT-3.5-turbo 모델을 활용하여 생성한 기사 데이터를 사용하였다. 향후 연구에서는 KULLM, KLUE-BERT와 같은 다양한 한국어 특화 언어모델을 활용하여 기사 데이터를 생성한 후 모델을 학습시켜 분류모델의 범용성을 더 높일 예정이다. 또한 활용 데이터의 양을 확장하여 기사의 제목 또한 분류 가능한 모델에 대해 연구할 예정이다.

<사사문구>

이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 "뉴스 기사 기계독해 데이터"를 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

참고문헌

[1] 전용철, A'I기자가 작성한 뉴스 기사가 뉴스 신뢰도에 미치는 영향:인간 기자와 AI기자의 비교 실험을 중심으로“, 한국콘텐츠학회논문지, 제24권, 제3호, pp168-pp179

[2] 변종국·남혜정, ““해리슨이 총에 맞았다”…머스크표 AI ‘그록’, 잘못된 기사 작성 논란“, 동아일보, 2024.07.18.

[3] 이국성, "Glove를 이용한 가짜 뉴스, 진짜 뉴스 기사 판별 개선에 대한 연구“, 석사학위논문, 남서울대학교, 2021

[4] 이동호 등 7명, "딥러닝 기법을 이용한 가짜뉴스 탐지“, 한국정보처리학회 학술대회논문집, 25권 1호, pp384-pp387

[5] 강주영·송민, "한국어 가짜 구매후기 생성과 탐지 성능 평가“, 지능정보연구 제30권 제2호, pp313-pp328

[6] KoBART 모델 관련 Github, https://github.com/SKT-AI/KoBART

[7] KoELECTRA 모델 관련 Github, https://github.com/monologg/KoELECTRA