TF-IDF와 KoBERT 모델을 이용한 인터넷 뉴스 신뢰도 판별

김나현[1], 서익원[2], 김정현[3], 손채영[4], 유동영[5]
[1234]홍익대학교 소프트웨어융합학과 학부생
[5]홍익대학교 소프트웨어융합학과 교수
knah0713@g.hongik.ac.kr, ikwon129@g.hongik.ac.kr, snu08029@g.hongik.ac.kr,
0_0ann@g.hongik.ac.kr, ydy@hongik.ac.kr

Identification of Internet news reliability using TF-IDF and KoBERT models

Na-Hyeon Kim[1], Ik-won Seo[2], Jeong-Hyeon Kim[3],
Chae-Young Son[4], Dong-Young Yoo[5]
[12345]Department of Software and Communication Engineering, Hongik University

요 약

디지털 환경이 진화함에 따라 가짜뉴스가 늘어나고 있다. 이를 판별하기 위해 법적 규제에 대한 논의가 있으나, 가짜뉴스에 대한 범위와 정의가 명확하지 않아 규제가 쉽지 않다. 본 논문에서는 이에 대한 대안으로 TF-IDF 기법과 KoBERT 모델을 이용한 키워드 추출 및 문장 유사도 분석을 통해 YouTube 플랫폼을 대상으로 한 가짜뉴스 판별을 위한 모델을 제안한다.

1. 서론

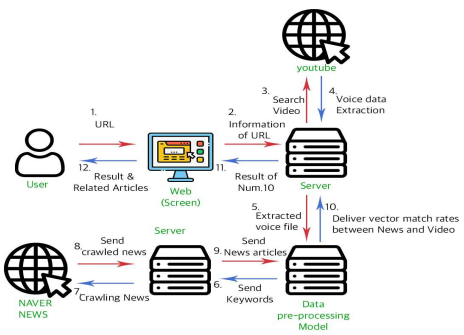
2022 언론수용자 조사에 따르면 인터넷을 통한 뉴스 이용률이 77.2%, 텔레비전 뉴스 이용률이 76.8%로 나타나며 이 중 20대와 30대의 인터넷 포털 뉴스 이용률은 평균 90.9%로 매우 높은 이용률을 보인다[6]. 이처럼 디지털 환경이 진화함에 따라 '가짜뉴스'의 위험성에 대응하기 위해 서울대학교 언론정보연구소에서 서비스 중인 SNU FactCheck와 같이 외부 기관 및 전문가와의 협업을 통하여 검증된 정보를 제공하는 방법과, 인공지능과 자연어 처리 기술을 활용하여 가짜뉴스를 판별하는 연구들이 대안으로 논의되고 있다[5][6]. 본 논문에서는 텍스트 마이닝을 활용한 YouTube 플랫폼 기반 영상 미디어의 가짜뉴스 판별법을 제안하고자 한다.

2. 관련 연구

인터넷상의 대용량 문서의 증가에 따라 주제 제시를 위한 키워드 추출 연구가 활발히 수행되고 있다. 키워드 추출을 위해 [7]에서는 단어들이 동시에 출현하는 통계적 정보를 활용하였으며, [8]에서는 인터넷 검색을 위한 색인 생성에 사용되는 PageRank 알고리즘을 이용하였다[1]. 문서 내 단어의 빈도수를 이용하여 문장 중요도를 결정하는 TF-IDF 기법도 있다. 하지만 이는 키워드 동시 발생 빈도를 활용하여 문장과 키워드의 중요도를 판단하기 때문에 구문 패턴 혹은 위치 등과 같은 문장의 관계를 정확히 이해할 수 없다는 단점이 존재한다.

반면 BERT 모델은 Google에서 공개된 자연어 처리를 위한 딥러닝 언어 모델로, 기존의 언어 모델과 달리 Transformer 모듈을 기반으로 양방향으로 입력 데이터의 맥락을 인코딩하여 문장 사이의 관계 학습 능력이 뛰어난 언어 모델이다[1]. 이를 활용하여 [4]에서는 BERT 모델을 이용하여 한국어 문맥 정보를 추출하는 시스템을 구현하였다.

3. 실험 설계 및 결과 예측



(그림 1) 시스템 구성도

관련 데이터 셋을 수집하기 위하여 2010년 1월 1일부터 2023년 12월 31일까지의 가짜뉴스 관련 기사를 수집할 것이다. 진위가 검증된 뉴스들은 언론사에서 제공하는 영상으로, 가짜뉴스들은 언론사 외의 YouTube 채널에 업로드된 영상들로 수집할 예정이다.

수집된 영상들의 링크는 Amazon에서 제공되는 “Amazon Transcribe” STT(Speech-to-text) 기술을 통해서 음성을 문자로 변환할 것이다.

<표 1> 형태소 분석기별 기능 비교[3]

	Kkma	Mecab	Hannanum	Komoran	KMA	홍익소
공백 삽입 기능	O	O	X	X	O	O
왜곡된 형태	△	X	△	O	△	△
신조어 처리	X	X	X	X	O	O
규칙 추가 기능	X	O	X	X	X	O

표 1은 KoNLPy 패키지에서 제공되는 네 가지 형태소 분석기(Kkma, Mecab, Hannanum, Komoran)와 KLT2010의 한국어 형태소 분석기(KMA)의 기능을 정리한 표이다. 본 연구는 유튜브 동영상을 대상으로 진행하므로 왜곡된 형태와 신조어 처리 기능을 지원하는 KMA를 사용한다[2]. 이에 더해 형태소 분석 시 입력 데이터가 주어질 때마다 필요한 형태소 분석 규칙을 추가한 ‘홍익소’ 형태소 분석기를 개발하여 고유명사 및 신조어 추출의 정확도를 높이고자 한다[3].

분절화된 문장들에 TF-IDF 기법을 적용하여 높은 빈도를 보이는 단어들을 키워드로 추출한 뒤, 키워드를 통해 웹 크롤링한 기사들과 KoBERT 모델을 이용하여 기사들 간 비교 분석을 통한 결과를 사용자에게 제공한다.

4. 결론

본 논문은 현대인들이 많이 사용하는 YouTube 동영상에서 추출한 텍스트의 데이터를 활용해 가짜뉴스를 판별하는 것을 목적으로 한다.

이전 연구에서 활용되던 자연어 처리기술(NLP)과 딥러닝 모델을 활용하여, 키워드 추출 및 유사도 분석을 수행하며, 이를 토대로 뉴스의 유사도를 분석하여 가짜뉴스를 식별하는 알고리즘을 개발하고자 한다. 또한, 다양한 데이터 셋을 활용하여 분석의 정확도와 신뢰도를 높이고, 텍스트 위주였던 분석 대상을 동영상 데이터의 텍스트 데이터 추출을 통한 범위의 확장에 의의를 둔다.

본 논문에서 제안하는 방법은 허위 정보의 확산이 빠르게 이루어지는 소셜 미디어에서도 유용하게 활용될 것으로 기대된다. 이를 통하여 사용자들의 올바른 정보 습득에 도움이 되며, 뉴스 제작자들의 윤리적인 책임 의식을 강화하는 데 기여할 것이라 예상한다.

참고문헌

[1] 이성직, 김한준, “TF IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법”, 한국전자거래학회지, Vol.14, No.4, pp.59-73, 2009.

[2] 원혜진, 이현영, 강승식, “대규모 텍스트 분석을 위한 형태소 분석기의 실행 성능 비교”, 한국컴퓨터종합학술대회, 한국정보과학회, 2020, pp.401-403.

[3] 유양우, 김현규, “응집도 점수 기반의 키워드 추출 정확도 향상을 위한 대화형 형태소 분석”, 한국컴퓨터정보학회논문지, Vol.25, No.12, pp.145-153, 2020.

[4] 유소엽, 정옥란, “BERT와 지식 그래프를 이용한 한국어 문맥 정보 추출 시스템”, Journal of Internet Coumputiong and Services, Vol.21, No.3, pp.123-131, 2020.

[5] 한유진 외 1, “검증 자료를 활용한 가짜뉴스 탐지 자동화 연구”, 정보처리학회논문지, 소프트웨어 및 데이터 공학, Vol.10, No.12, 2021.

[6] 오세욱 외 3명, "2022 언론수용자 조사", 한국언론진흥재단, 2022.

[7] Matsuo and 3, “Keyword extraction from a single document using word co-occurrence statistical information”, International Journal on Artificial Intelligence Tools, Vol.13, No.1, 2003, pp.157-169.

[8] Wang and 5, “Keyword extraction based on PageRank”, Lecture notes in computer science, 2007, pp.857-864.