Bi-LSTM과 토픽모델링을 활용한 카카오톡, 인터넷 가짜뉴스 판별 서비스

심국보*, 이승호*, 정준호*, 이기영**
*인천대학교 정보통신공학과
**인천대학교 사물인터넷 빅데이터 연구센터
Sim0494@naver.com, dltmdgh@gmail.com, jjh3706@gmail.com, kylee@inu.ac.kr

Kakao Talk, Internet fake news identification service using Bi-LSTM and topic modeling

Kuk-Bo Shim*, Seung-Ho Lee, Jun-Ho Jeong, Ki-Young Lee
*Dept. of Info and Telecom Eng, Incheon National University
** IoT and Big-Data Research Center, Dept. of Info and Telecom Eng.
Incheon National University

요 약

현재 영어 기반의 기술 팩트체크 서비스는 다양하지만 한국 기반 팩트체크 서비스는 비기술적(언론인 등 전문가의 교차 검증을 통한 팩트체크)이 주를 이루고 있으며, 기술 팩트체크 서비스가 많이 시행되지 않고 있다. . 본 논문에서는 기술적인 요소와 비기술적인 요소의 서비스를 함께 사용할 때 허위 정보를 가장 정확하게 식별할 수 있기 때문에 한국어 기반의 자연어 처리 기술을 이용한 팩트체킹 서비스를 제안한다.

Ⅰ. 서론

글로벌 조사 전문기관인 'Ipsos'에서 전세계 25 개국을 대상으로 '가짜 뉴스'에 대해 조사한 결과, 한국의 경우 85%의 국민이 가짜 뉴스에 속은 경험이 있는 것으로 나타났다. 특히 장,노년층의 경우, 노화 및 치매 등으로 인해 디지털 정보격차가 점점 더 커져 정보가 올바른지 판단하기에도 큰 어려움이 있으며 이는 세대 간의 갈등을 유발하고 불필요한 경제적, 사회적 피해를 야기한다[1]. 정형화된 뉴스기사가 아닌 SNS 를 통해 유포되는 비정형화된 가짜뉴스, 가짜정보 또한 매우 많은데, 이를 검증하는 서비스가 구현되어 있지 않아 큰 어려움을 겪고 있다.

본 논문에서는 머신러닝 기법이 적용된 자연어 처리 알고리즘을 활용한 가짜 뉴스, 가짜 정보 판독 어플리케이션을 개발하여 장,노년층을 비롯한 일반 사용자들이 쉽게 가짜 뉴스를 판단할 수 있도록 함으로써 이를 통해 잘못된 정보로부터 사회를 보호하고, 수많은 데이터 속에서 올바른 정보를 추출하는 시각을 함양하고자 한다.

Ⅱ. 이론적 배경

2.1 토픽 모델링

토픽 모델링(Topic Modeling)은 기계 학습 및 자연어 처리 분야에서 토픽이라는 문서 집합의 추상적인 주제를 발견하기 위한 통계적 모델 중 하나로, 텍스트 본문의 숨겨진 의미 구조를 발견하기 위해 사용되는 텍스트 마이닝 기법이다. 각 문서에 포함된 용어의 빈도수에 기반하여 유사 문사를 그룹화한 뒤 각 그룹을 대표하는 주요 용어들을 추출함으로써 해당 그룹의 토픽 키워드 집합을 제시하는 방식으로 이루어지며 이 때 사용되는 문서는 문서, 제목, 요약, 본문, 댓글 등을 포함하는 넓은 개념을 의미한다[2].

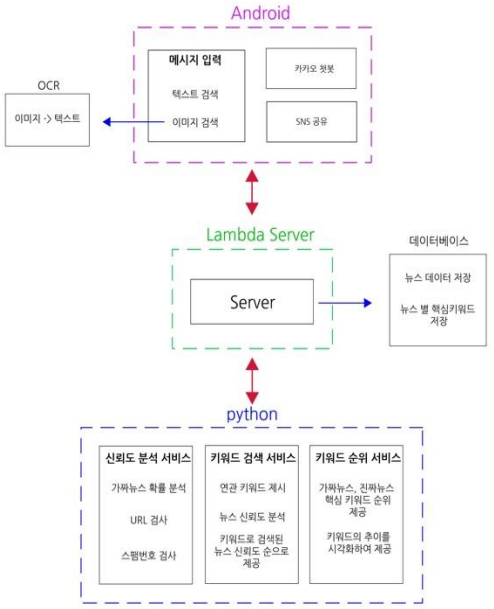
2.2 Bi-LSTM

게이트 기법을 통해 순환 신경 회로망(RNN)의 한계를 극복한 모델인 LSTM을 순방향 뿐 아니라 역방향의 결과를 함께 이용하는 모델이다. 문맥(Context)을 기반으로 하는 연관성 분석에 유리하여 속도를 고려한 다양한 NLP 문제에 널리 활용되고 있다[3].

본 서비스는 한글 형태의 카카오톡 가짜 정보와 가짜 뉴스로 알려진 언론사의 기사들을 수집하여 CSV 파일 형식의 DB 를 구축하고, 해당 데이터에 대하여 형태소 분석 및 토픽 모델링을 수행한 후 Bi-LSTM 알고리즘 을 활용하여 패턴 분석을 통한 가짜 뉴스의 패턴 정보 추출 및 저장을 수행한다. 또한 AWS 를 활용하여 카카오톡 가짜 정보 및 가짜 뉴스 데이터 수집을 자동화하여, 지속적으로 생성되는 새로운 내용의 카카오톡 가짜 정보 및 가짜 뉴스에 대한 신뢰도를 수치로 나타냄으로써, 가짜 정보 및 가짜 뉴스로 인한 개인의 정신적 피해, 금전적 피해 및 집단의 경제적 피해, 사회적 피해를 예방한다.

Ⅲ. 구현

본 서비스를 개발하기 위한 주요 IDE 로는 파이썬(Python) 3.8 을 사용하기 위한 Pycharm 및 Anaconda 를 활용하며, 가짜 뉴스 및 기사에 대한 데이터베이스 구축을 위해 AWS 의 S3 와 RDS 를 활용한다. 또한 사용자의 편의성을 최대한으로 높이기 위해, 본 서비스의 형태는 안드로이드 어플리케이션을 채택하며, 이에 따라 Android Studio 를 활용한다.
본 어플리케이션의 경우 기본적으로 백그라운드에서 동작하며, 사용자가 의문을 갖는 카카오톡 가짜 정보 및 기사를 텍스트 형식으로 입력(Input)을 받는다. 그림 1 은 위 구현방법을 간단히 구성도로 나타낸 것이다.



(그림 1) SW 구성도

텍스트를 입력 받으면 Lambda 서버에서 데이터와 핵심키워드를 저장 후, python 서버를 통해 확률을 분석한다. 분석한 결과를 다시 Lambda 서버로 전달 후 사용자에게 보여준다.

3.1 데이터 수집

뉴스 데이터의 경우 공신력이 있다고 판단되는 'SNU FactCheck' 사이트를 중점으로 가짜 뉴스 혹은 진짜 뉴스로 판별된 뉴스기사를 수집하였고, 가짜 뉴스 판별 정확도를 향상시키기 위해 진짜, 가짜 뉴스에 대한 네이버 뉴스 댓글을 수집하여 학습 데이터로 사용하였다. 카카오톡의 경우 유포되고 있는 가짜 뉴스의 객관성 정도를 판별하기 위해 15 만건의 리뷰 데이터를 수집하여 학습 데이터로 사용하였다.

3.2 데이터 처리

수집한 데이터를 카카오톡 가짜 뉴스와, 인터넷 뉴스 기사 두 가지로 나누어 분석한다.

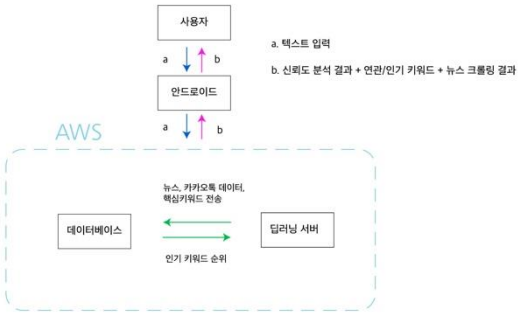
3.2.1 인터넷 뉴스 기사

분석하고자 하는 뉴스 기사를 검색하면 TF-IDF 알고리즘으로 토픽 모델링을 수행하여 해당 내용과 관련된 뉴스 데이터를 조회한다. 먼저, 토픽모델링으로 추출한 핵심 키워드로 자체적으로 수집한 SNU 팩트체크 뉴스 데이터를 검색하고, 네이버 뉴스 크롤링을 진행한다. 검색된 모든 네이버 뉴스에 대하여 팩트체크 뉴스 데이터로 학습한 Bi-LSTM 모델과 네이버 댓글 반응으로 학습한 Bi-LSTM 모델에 입력하여 각 뉴스의 신뢰도를 분석한다.

3.2.2 카카오톡 가짜 뉴스

분석하고자 하는 카카오톡 텍스트를 검색하면 인터넷 기사 검색과 마찬가지로, 토픽 모델링을 수행하여 해당 내용과 관련된 뉴스 데이터를 조회 및 분석한다. 이에 더하여 해당 내용을 리뷰 데이터로 학습된 감성 분석 LSTM 모델에 입력하여 객관성 정도를 판단한다.

3.3 안드로이드 연동



(그림 2) 하드웨어 구성도

그림 2 는 하드웨어 구성도를 나타내며, 사용자가 높은 편의성을 가지고 사용할 수 있도록 UI/UX 를 고려하여 Android Studio 를 활용하여 어플리케이션을 구현한 후, AWS EC2 상에서 작동되도록 한 본 연구의 Logic 의 결과를 AWS EC2 와 통신하여 어플리케이션 상에 출력하도록 한다.

Ⅳ. 구현 결과

4.1 가짜 뉴스 판별 모델

4.1.1 under sampling (데이터 불균형)

SNU 팩트체크 뉴스 데이터를 수집하여 이를 학습 데이터로 가짜 뉴스 판별 모델을 구현하였다. 하지만 수집한 뉴스 데이터의 경우 가짜뉴스 80%, 진짜 뉴스 20%의 비율로 구성되어 있어 데이터 불균형으로 인한 과적합 문제가 발생한다. 이를 해결하기 위해 가짜 뉴스의 비율을 진짜 뉴스의 비율에 맞춰주는 under sampling을 진행하였다.

4.1.2 데이터 분류

과적합 문제는 해결하였지만 여전히 가짜뉴스를 판별하기에는 모델의 정확도가 부족하다고 판단하여 전체 뉴스 데이터를 각 분야(정치, 경제, 사회, 기타)별로 분리하여 학습 데이터를 재구성하였다.

4.1.3 댓글 반응 분석

또한 팩트체크 된 뉴스의 수가 부족하고, 뉴스 데이터만으로는 새로운 뉴스를 판별하기 어렵기 때문에 뉴스 데이터 분석에 더하여 해당 뉴스에 대한 네이버 댓글 반응을 분석하여 기존 가짜 뉴스 판별 모델을 보완한다.

4.2 실험 결과

가짜 뉴스 판별 모델을 under sampling, category, comments 기준으로 표1과 같이 분류한다.

<표 1> 가짜 뉴스 판별 모델 비교

	정치	경제	사회	기타
카테고리 (under sampling) with Comments	53.2%	41.9%	64.5%	63.0%
카테고리 (under sampling)	53.2%	48.3%	60.7%	67.3%
전체 (under sampling)	50.6%	54.8%	55.6%	54.3%
전체	67.5%	64.5%	68.3%	67.2%

under sampling된 모델이 데이터 불균형 문제를 해결하지 않은 전체 데이터로 학습한 모델과 비교하여 정확도가 낮게 나타난다. 이는 현재 수집한 팩트체크 뉴스의 수가 매우 부족하고, under sampling으로 인해 전체 학습 데이터의 60%가 삭제되었기 때문에 발생한 데이터 부족 현상으로 보여진다. 실제로 전체 데이터 중 다른 카테고리와 비교하여 매우 낮은 비율을 차지하는 경제 카테고리 모델의 경우 학습된 데이터의 수가 충분하지 않아 정확도가 많이 낮아진 것을 확인할 수 있다.

Ⅴ. 결론 및 향후 연구 방향

본 연구에서는 카카오톡 가짜 정보에 대한 진실,거짓 여부를 객관적으로 판단하기에는 어려움이 있다고 판단되어, 토픽모델링을 통한 키워드 추출과 Bi-LSTM 을 적용한 감성분석으로 해당 가짜 정보 내용에 대하여 주관적/객관적인 정도를 수치로 나타내는 데에 주목하였다. 인터넷 기사의 경우, 토픽모델링으로 키워드 추출 및 해당 키워드와 연관 있는 검증된 기사와 네이버 뉴스 댓글을 활용한다. 하지만 현재 제 4.2 절에서 확인된 바와 같이 팩트체크 된 뉴스 데이터가 많이 부족하여 뉴스 기사와 댓글 반응에 대한 분석 정확도가 낮기 때문에 추후에는 가짜뉴스 DB 공유 웹사이트를 제작하여 자체적으로 수집한 가짜뉴스 데이터를 제공하고, 사용자가 직접 가짜뉴스를 제보할 수 있도록 할 예정이다. 이를 통해 분석 알고리즘의 성능을 향상시키고, DB 를 공개하여 가짜뉴스 관련 연구나 비즈니스에 활용할 수 있다.

ACKNOWLEDGEMENT

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] Ipsos, "'진짜'일까 '가짜'일까, 페이크 뉴스(가짜뉴스 fake news)에 대한 국내외 의견", 2018, https://www.ipsos.com/ko-kr/ibsoseu-peobeullig-jinjjailkka-gajjailkka-peikeu-nyuseugajjanyuseu-fake-newse-daehan-gugnaeoe

[2] 현윤진, 김남규. (2018). 뉴스와 소셜 데이터를. 활용한 텍스트 기반 가짜 뉴스 탐지 방법론. 한국전자거래학회지, 23(4), pp.19-39.

[3] 박해선, 혼자 공부하는 머신러닝 + 딥러닝, 한빛. 미디어, 2020.

[4] 벤자민 존스턴, Applied Unsupervised Learning with Python, 에이콘, 2019.

[5] 좌희정, 오동석, 임희석, 2019. 자동화기반의 가짜. 뉴스 탐지를 위한 연구 분석, 한국융합학회논문지 제 10권 제 7호 (4).

[6] 홍태석, 강상우, 서정연, 2018. Fake News. Detection Using Deep Neural Network 한국정보과학회, 2018 한국소프트웨어 종합 학술 대회 논문집.