딥러닝 전이학습을 이용한 경량 트렌드 분석 시스템 설계 및 구현

신종호* · 안수빈 · 박태영 · 방승철 · 노기섭

청주대학교

Design and implementation of trend analysis system through deep learning transfer learning

Jongho Shin* · Suvin An · Taeyoung Park · Seungcheol Bang · Giseop Noh

CheongJu University

E-mail : {sjh0804_22 / ssubin0405 / mu07010 / bbt1250 / kafa46}@cju.ac.kr

요 약

최근 코로나로 인해 집에 있는 시간이 많아진 소비자들이 증가함에 따라 비대면으로 쉽게 사용 할 수 있는 SNS와 OTT등 디지털 소비를 하는 시간이 자연스럽게 늘어났다. 코로나가 발생한 2019년 이후 디지털 소비는 44%에서 82%로 두 배가량 증가하였고 트렌드가 빠르게 변화하는 디지털 특성상 소비자들의 감성을 분석하여 트렌드를 신속, 정확하게 파악하여 적용하는 것은 중요하다. 그러나 대기업 수준의 시스템이 아닌 소규모 시스템에서 감성분석을 활용한 서비스를 실제로 구현하기에는 제약 사항이 있으며 실제 서비스 되는 경우도 많지 않다. 하지만 소규모 시스템이라도 간편하게 소비자들 트렌드 분석을 할 수 있다면 빠르게 변화하는 현대사회에 도움이 될 것이다. 본 논문에서는 BERT Model의 Transfer Learning(Fine Tuning)을 통해 학습 네트워크를 구축하고, 실시간 데이터 수집을 위한 Crawler를 연동하는 경량 트렌드 분석 시스템을 제안한다.

ABSTRACT

Recently, as more consumers spend more time at home due to COVID-19, the time spent on digital consumption such as SNS and OTT, which can be easily used non-face-to-face, naturally increased. Since 2019, when COVID-19 occurred, digital consumption has doubled from 44% to 82%, and it is important to quickly and accurately grasp and apply trends by analyzing consumers' emotions due to the rapidly changing digital characteristics. However, there are limitations in actually implementing services using emotional analysis in small systems rather than large-scale systems, and there are not many cases where they are actually serviced. However, if even a small system can easily analyze consumer trends, it will help the rapidly changing modern society. In this paper, we propose a lightweight trend analysis system that builds a learning network through Transfer Learning (Fine Tuning) of the BERT Model and interlocks Crawler for real-time data collection.

키워드

Transfer Learning; Fine tuning; BERT Model; Multiprocessing Crawler;

Ⅰ. 서 론

코로나19에 대응하기 위한 사회적 거리두기는 사람들이 비대면 서비스를 선호하게 했다. 한국소비자원에서 조사한 ‘2021 한국의 소비생활 지표’에 따르면 코로나 이후 디지털 소비는 44%에서 82.1%로 2배가량 증가하였으며 유형별로 인터넷·모바일 쇼핑(13.0%), TV홈쇼핑(22.6%), SNS 플랫폼(16.7%) 각각 증가하였다**[1]**. 이에 맞추어 기업과 공공기관 등은 핵심 서비스들을 비대면으로 제공하기 위해 기존의 오프라인 서비스들을 온라인망 중심의 시스템으로 전환하였고**[2]**, 사회 전반의 변화를

* speaker author

현재와 과거의 데이터에 근거해 선제적으로 대응하기 위한 트렌드 분석의 관심은 높아졌다. 하지만 기존의 트렌드 분석 방법인 설문조사, 조사기관, 데이터 분석 사이트 등과 같은 방식들은 금전적인 지출과 시간이 소요가 불가피하였다. 그 중 트렌드 분석 사이트의 경우 단순히 키워드에 대한 빅데이터 결과들을 기간별로 보여주거나 추천해주는 정도에 그쳤을 뿐 사용자들의 감정을 분석하는 시스템은 적었고 소규모 시스템에서 구동할 수 있는 트렌드 분석 시스템은 존재하지 않았다.

본 논문에서는 소규모 시스템으로 딥러닝 전이학습을 통한 감성분석 트렌드 분석이 가능함을 확인하고 Web Application으로 이용할 수 있는 경량 트렌드 분석 시스템을 제안한다.

Ⅱ. 시스템 설계 및 개발

1) 시스템구성 및 동작순서

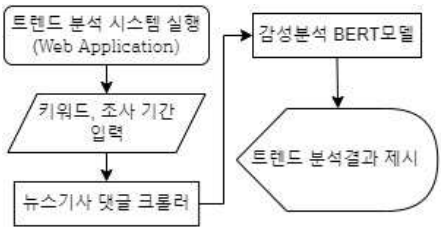

그림 1. 시스템 구성 및 순서

시스템은 데이터 수집을 위한 크롤러, 감성분석 언어모델, Web Application으로 서비스 되는 응용 시스템 3가지로 구성되어 있다.

시스템의 동작은 분석을 원하는 키워드와 조사기간을 웹사이트에 전달하면 크롤러가 해당 입력값에 일치하는 데이터를 수집하고 BERT 모델의 감성분석을 통해 트렌드 파악에 도움을 주는 정보들을 웹사이트로 제공하게 된다.

2) 데이터 수집

본 논문에서는 감성분석을 위해 사람들의 여론이 가장 직설적이고 활발한 네이버 뉴스 댓글을 크롤링하였다. 뉴스 댓글 수집을 위한 크롤러 내부 동작 과정은 그림 2와 같고 JavaScript가 동적으로 만든 데이터를 크롤링하기 위해 Selenium을 사용하였다.

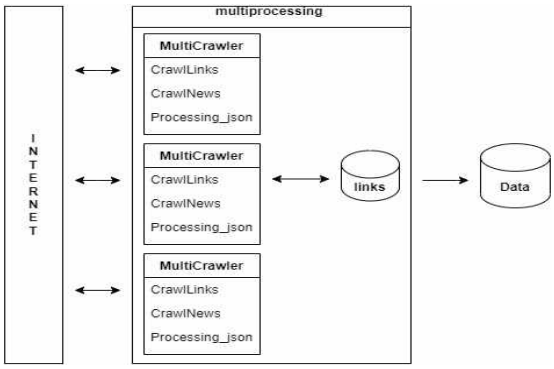

그림 2. 클래스 다이어그램

사용자가 크롤링 키워드에 대한 옵션을 지정하면 Web Application 서버 안의 MultiCrawler 클래스의 CrawlLinks 매소드로 옵션값들을 보내준다. CrawlLinks 에서는 받은 옵션값과 일치하는 네이버 뉴스 기사의 링크들을 기간별로 저장한 다음 CrawlNews 매소드에서 저장된 링크를 통해 뉴스기사에 접속하여 모든 댓글들을 크롤링한다. 마지막으로 Processing_json 매소드를 통해 수집한 댓글 내용에서 불필요한 이모티콘과 특수문자들을 제거하는 전처리 과정을 거친다.

위 모든 과정은 파이썬의 multiprocessing 모듈을 사용하여 병렬화 작업이 진행되며 보다 빠르게 뉴스 댓글, 작성 날짜 데이터들을 수집할 수 있다.

3) 딥러닝 모델 구축

본 논문에서는 감성분석을 이용한 트렌드 분석을 위해 BERT Model에 Transfer Learning을 통한 Fine-Tuning을 진행하였다.

맨 처음, Fine-Tuning에 사용할 데이터셋을 불러온다. 학습을 위한 데이터는 AI HUB 내의 '감성대화 말뭉치', '한국어 SNS대화 데이터', '한국어 대화 요약' 약 70만 개의 데이터를 전처리 후 사용한다. 데이터를 토큰화 하기 전, 해당 텍스트 데이터를 사용할 용도에 맞게 정제, 및 정규화 과정을 거친다. BERT Model 내에 있는 토크나이저를 통해 토크나이징을 진행하였고, 이 데이터에 어텐션 마스크를 입력했다. 이 모델의 경우 훈련 데이터와 테스트 데이터를 8 : 2로 나누어 진행했다. 데이터 로더를 통해 학습에 필요한 데이터 구조를 생성하고 학습시 배치 사이즈만큼 데이터를 가져온다. 사전 학습된 BERT Model의 옵티마이저는 AdamW를 사용하였고 학습률은 1e-5로 설정한다. 에폭은 4로 설정하여 전체 데이터 셋을 4번 학습 시켰다.

학습 결과 문장에 따라 감성분석을 통해 문장이 긍정적인 의미인지 부정적인 의미인지 구별 가능한 경량 트렌드 분석 딥러닝 모델을 구축하였다.

4) 응용시스템 구현

응용시스템은 Web Application으로 경량 시스템의 목적에 맞게 다른 Web Framework보다 가벼운 파이썬의 Flask를 통해 구현하였다.

웹사이트는 분석할 키워드 및 조사 기간을 설정하는 검색 입력 페이지와 결과를 보여주는 결과 페이지, 서버에 저장된 모든 트렌드 분석 결과를 볼 수 있는 페이지가 있다. 결과 페이지에서는 검색 키워드의 기간별 긍·부정 그래프와 관심도 그래프, 최빈 단어 상위 50개 워드 클라우드, 키워드의 총 긍·부정 비율을 원그래프로 확인 가능하다.

Ⅲ. 시스템 테스트 및 평가

<표3> '윤석열 국정평가' 트렌드 분석결과 비교

구분	2022년 8월 1주차 주간 집계				
	7월 1주	*7월 2주*	*7월 3주*	*7월 4주*	*8월 1주*
긍정 (리얼미터)	37.0%	33.4%	33.3%	33.1%	29.3%
부정 (리얼미터)	57.0%	63.3%	63.4%	64.5%	67.8%
긍정 (경량 시스템)	42.2%	38.7%	36.8%	37.3%	33.8%
부정 (경량 시스템)	57.8%	61.3%	63.2%	62.7%	66.2%

경량 트렌드 분석 시스템의 실용성과 정확도를 확인하기 위해 여론 조사 매체 '리얼미터'에서 발표한 '윤석열 국정평가' 1주차 여론 조사와 본 시스템이 분석한 여론 조사를 비교하였다**[6]**. 결과는 <표3>과 같다.

<표3>과 같이 설문조사를 통한 키워드 '윤석열 국정평가'에 대한 여론 평가 결과가 경량 트렌드 분석 시스템의 결과와 긍·부정 비율이 같음을 확인할 수 있다. 이를 통해 소규모 시스템으로도 데이터 수집과 인공지능 감성분석을 통한 트렌드 분석 방식이 가능하고 분석 정확도와 실용성이 기존의 조사 방법에 뒤처지지 않는 선에서 시간도 절약됨을 확인 할 수 있었다.

Ⅳ. 결 론

본 논문에서는 소규모 시스템으로도 구동 가능한 경량 트렌드 분석 시스템을 제안했다. 분석 시스템은 크롤링을 통해 실시간으로 데이터를 수집을하고 BERT 모델에 감성분석을 위한 Fine-Tuning을 진행하여 트렌드 분석이 가능하다.

제안한 경량 시스템을 구현하여 작동한 결과 여론 조사기관인 리얼미터에서 발표한 '윤석열 국정평가' 분석 결과와 최대 오차가 5.3%로 큰 차이가 없었다, 따라서 경량 시스템으로도 딥러닝을 확용한 트렌드 분석 시스템이 구현가능하고 정확도에서도 기존의 분석 방식와 차이가 없어 보다 빠르고 편리하게 트렌드 분석이 가능함을 확인하였다.

Acknowledgement

The authors thank Junjun Zhang for advice on BERT model and Fine-Tuning

References

[1] M.Y. Heo, and B.K. Lim, "A Study on the Direction of Consumer Policy in the Acceleration of Digital Transformation after COVID-19", Policy Research 21-01, Korea Consumer Agency, 2021.

[2] D.H. Oh, "[Digital transformation has been accelerated due to the spread of non-face-to-face culture] Digital transformation through the spread of non-face-to-face culture after COVID-19, highlighting the core of urban competitiveness," Busan Development Forum, Vol. 192, pp. 68-73, December. 2021.

[3] Chi Hoon Lee, Yeon Ji Lee, and Donghee Lee, "A Study of Fine Tuning Pre-Trained Korean BERT for Question Answering Performance Development," *Journal of Information Technology Services*, Vol. 19, No. 5, pp. 83-91, 2020.

[4] H. C. Kim and S. H. Chae. "Design and Implementation of a High Performance Web Crawler," *Journal of Digital Contents Society*, Vol. 4, No. 2, pp. 127-137, December. 2003.

[5] H.S. Kim, N. Han, and S.J. Lim, "Web Crawler Service Implementation for Information Retrieval based on Big Data Analysis," *Journal of Digital Contents Society*, Vol. 18, No. 5, pp. 933-942, Aug. 2017.

[6] Trends in the week of August 1st week of Realmeter [Internet]. Available: http://www.realmeter.net/dfu0u0q28uf/.