KoBERT를 활용한 식품 게시글 카테고리 분류 모델의 설계†

현태민[1], 김희진[1], 임은지[1], 길준민[2,*]
[1]대구가톨릭대학교 컴퓨터공학전공
[2]대구가톨릭대학교 컴퓨터소프트웨어학부
taemin14215@cu.ac.kr, dntwk@cu.ac.kr, dmswl08087@cu.ac.kr, jmgil@cu.ac.kr

Design of Category Classification Model for Food Posts using KoBERT

Tae Min Hyeon[1], Hui Jin Kim[1], Eun Zi Lim[1], Joon-Min Gil[2,*]
[1]Major in Computer Engineering, Daegu Catholic University
[2]School of Computer Software Engineering, Daegu Catholic University

요 약

본 논문에서는 식품 판매 게시글에 대한 카테고리 분류를 위해 자연어처리 모델인 KoBERT 모델에 기반하여 식품 판매글에 대한 카테고리 분류 모델을 설계하고 구현한다. 본 논문을 통해 구현된 식품 판매 게시글의 카테고리 분류 모델은 정확도 평가에 대해서 비교적 우수한 성능을 산출하였다.

1. 서론

2021년 기준 1인 가구는 전체 가구의 33.4%인 716만 6천 가구로 가장 큰 비중을 차지하는 가구의 형태로 사회에 자리 잡았다[1]. 이러한 1인 가구는 매 끼니별 식사 해결을 위해 배달 및 포장 등의 간편식을 이용하는 비중이 점점 늘어나고 있다. 일주일 평균 1.6회를 이용하는 것으로 보고되었다. [2]. 배달 음식들은 영양소가 고르지 않으며, 기름진 경우가 많아 영양 불균형과 비만 문제를 야기한다. 건강 문제뿐만 아니라 배달 및 포장 시 일회용기가 주로 사용되며, 배달 시 최소 금액을 채우기 위해 필요 이상의 음식을 주문하게 되어 음식물 쓰레기를 발생시킨다. 또한 간편식에 사용되는 일회용기가 1인 가구의 쓰레기 배출량을 증폭시킨다. 1인 가구의 식습관으로 발생하는 문제는 사회 전반적인 문제로 언급되고 있다.

이러한 문제를 해결하기 위해서는 1인 가구가 직접 요리하여 식사를 해결하도록 해야 한다. 현재 시장에 유통되는 식자재들은 유통기한이 짧고, 1인 가구가 소비하기 많은 양으로 소분되어 소비자들이 구매를 꺼린다. 별도로 1인 가구를 위한 식재료 유통이 필요한 실상이다.

그러므로 1인 가구를 대상으로 필요한 양의 식자재를 구매하고, 많은 양으로 포장된 식자재인 경우 공동으로 구매하거나 판매할 수 있도록 하는 플랫폼을 개발한다.

이러한 목적과 더불어 본 논문에서는 이용자가 플랫폼을 편리하게 사용할 수 있도록 자연어 처리 기술인 KoBERT 모델을 이용한 카테고리 추천 기능을 구현하고자 한다.

2. 식품 판매 게시글에 따른 카테고리 분류

본 논문의 식품에 따른 카테고리 분류 모델은 식품 판매글 제목 데이터를 KoBERT 모델에 적용하는 과정으로 수행된다. 이를 위한 설계 및 구현을 위해 식품 판매글의 제목을 기반으로 카테고리 분석을 수행하는데, 먼저 식품 판매글의 제목에 대한 크롤링 방법을 기술한다. 다음으로 식품에 따른 카테고리 분류를 위한 KoBERT 모델에 기반한 학습 방법과 카테고리 예측 모델에 의한 카테고리 분류 방법에 대해서 구체적으로 살펴본다.

2.1 크롤링

본 논문에서는 네이버 쇼핑몰[3]에서 각 식품에 대한 총 27,303개의 판매글 제목 데이터를 웹크롤링을 통해 얻어내었다. <표 1>은 본 논문의 크롤링을 통해 생성된 식품 판매글 제목 데이터의 구성을 보여준다.

† 본 연구는 과학기술정보통신부 및 정보통신기획평가원에서 주관하여 진행하는 'SW중심대학사업'의 결과물입니다(2019-0-01056).

* 교신저자(Corresponding Author)

<표 1> 식품 판매 게시글 제목 데이터 구성

항목	게시글 개수
채소	2073
과일	6067
정육	940
빵/떡	2133
과자/음료/간식	3211
양념/오일/조미료/가루/견과류	2843
유제품/계란	2209
수산/건어물	2206
면/통조림/가공식품	2838
커피/원두/차	898
반찬	946
담배/주류	939

2.2 KoBERT를 활용한 카테고리 분류

KoBERT는 구글 BERT의 한국어 성능 한계를 해결하기 위해 SKTBrain에서 개발한 모델이다[4].

본 논문에서는 기존에 구축된 KoBERT 언어모델에 식품 판매글 제목 데이터를 추가로 학습하는 방식인 미세 조정(fine tunning)을 적용하였다. 본 논문에서는 미세 조정된 KoBERT 언어모델을 사용하여 게시글에 따른 식품 카테고리 분류를 수행한다.

3. 실험

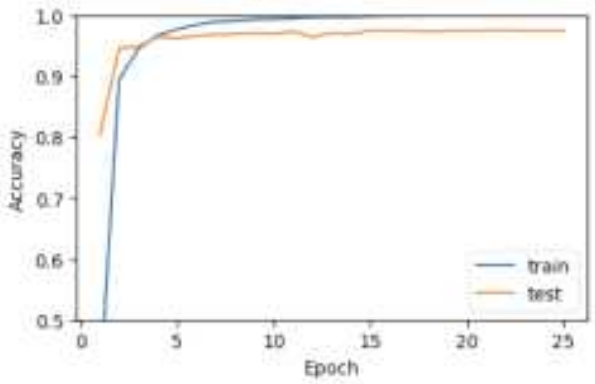
3.1 실험 환경

먼저 게시글 분류에 적합한 KoBERT 언어모델의 생성을 위해 앞서 <표 1>에서 제시된 데이터가 사용되었다. 총 27,303개의 판매 게시글 제목 데이터에 대해서 8:2의 비율로 학습 데이터와 테스트 데이터로 분리하여 사용하였다. 학습을 위한 배치 크기(batch size)는 64로 설정하였으며, 학습률(learning rate)은 $3\times e^{-5}$로 설정하였다. 총 25번의 학습 반복(epoch)을 통해 학습을 진행하였으며, 학습 수행은 구글 코랩 환경에서 GPU 장치를 이용하였다.

3.2 실험 결과

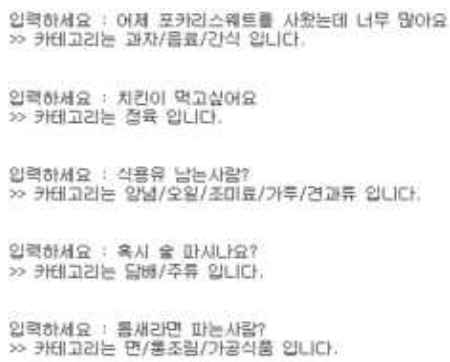
본 논문의 식품에 따른 카테고리 분류 모델의 성능은 학습 반복 횟수에 정확도(accuracy)로 측정하였다. 여기서 정확도는 실제값과 예측값을 비교하여 올바르게 예측한 비율을 의미한다.

학습 데이터와 테스트 데이터를 나누어 정확도를 측정하였으며, (그림 1)은 반복 횟수에 따른 정확도를 나타낸다. (그림 1)의 결과는 학습 데이터에 대해서는 1.0의 값으로 수렴하며 테스트 데이터에 대해서는 0.96에서 0.97 사이의 값을 가짐을 보여준다.



(그림 1) 학습 반복 횟수에 따른 정확도 성능

또한 (그림 2)와 같이 생성된 모델의 입력으로 식품에 관련된 글을 제시하였을 때 사용자가 원하는 결과로 분류하는 것을 확인할 수 있다.



(그림 2) 식품 게시글 카테고리 분류 모델 사용 예시

4. 결론

본 논문에서는 KoBERT를 이용한 자연어 처리 기술을 활용하여 게시글에 따른 식품 카테고리 분류 모델을 설계하고 구현하였다. 식품 판매글의 제목을 크롤링하여 데이터를 수집하였고, KoBERT 모델을 활용하여 식품 판매글의 제목에 대한 카테고리 분석을 수행하였다. 성능 평가 결과 약 97% 정도의 높은 정확도 성능을 보여주었다.

본 모델을 이용하여 식품 판매글에 대한 카테고리 분류 결과를 도출하였으나, 제안된 방식을 식품 분야뿐 아니라 다양한 분야에 적용할 수 있을 것으로 기대한다.

참고문헌

[1] 통계청, 2022 통계로 보는 1인가구, 2022년 12월
[2] Opensurvey, 1인 가구 트렌드 리포트, 2022년 11월
[3] 네이버 쇼핑몰, https://shopping.naver.com/home
[4] KoBERT, https://github.com/SKTBrain/KoBERT