KoBERT 기반 웹 확장 프로그램을 이용한 블로그 리뷰의 광고성 판별 시스템

이상현[O], 조대수[*], 김요한[*]
[O]동서대학교 소프트웨어학과,
[*]동서대학교 소프트웨어학과
e-mail: hyeon012366@gmail.com[O], dscho@dongseo.ac.kr[*], yhkim@dongseo.ac.kr[*]

A System for Determining the Advertising Property of Blog Reviews using a KoBERT-based Web Extension Program

Sang-Hyeon Lee[O], Dae-Soo Cho[*], Yo-Han Kim[*]
[O]Dept. of Software, Dongseo University,
[*]Dept. of Software, Dongseo University

● 요 약 ●

온라인 리뷰는 소비자에게 중요한 정보원이지만, 광고성 리뷰는 왜곡된 정보를 제공할 수 있다. 본 연구는 블로그 리뷰 텍스트를 분석하여 광고성 여부를 판단하고 확률값을 제공하는 확장 프로그램을 제안한다. 이를 위해 네이버 블로그 리뷰 데이터를 수집하고 라벨링한 후, KoBERT 모델을 학습시켰다. 실험 결과, 제안 시스템은 높은 정확도로 광고성 리뷰를 식별하고, 확률 예측 기능을 통해 신뢰성 있는 정보를 제공할 수 있었다. 또한 이시스템을 웹 브라우저 확장 프로그램으로 구현하여 사용자가 실시간으로 리뷰의 광고성 여부와 신뢰도를 확인할 수 있도록 하였다.

키워드: 광고성 리뷰(Advertising Review) 웹 확장 프로그램(Web Extensions), KoBERT, 자연어 처리(Natural Language Processing), OCR(Optical Character Recognition), 크롤링(Web Crawling)

I. Introduction

온라인 리뷰는 소비자에게 중요한 구매 결정 정보를 제공하지만, 일부는 광고 목적을 가진 리뷰일 수 있다. 이러한 광고성 리뷰는 소비자에게 왜곡된 정보를 전달해 올바른 구매 결정을 방해할 수 있다. 따라서 광고성 리뷰를 식별하고 분석하는 기술에 대해 연구가 필요하다. KoBERT 모델은 한국어 텍스트 처리에 특화된 BERT 모델로, 한국어의 고유한 언어적 특성과 문맥을 효과적으로 이해하고 분석할 수 있다. 본 연구에서는 KoBERT 모델을 기반으로 블로그 리뷰의 광고성 여부를 판별하는 시스템을 제안하고, 이를 웹 브라우저 확장 프로그램으로 구현하였다.

II. Related Work

블로그 및 SNS에서 리뷰와 광고성 글을 판별하는 다양한 방법들이 제시되어 왔다. 시계열 데이터를 학습하기 위한 LSTM-Attention 모델을 이용한 광고성 리뷰 분류 방법은 웹 크롤링으로 수집한 블로그 본문을 특정 문구로 라벨링하여 광고성 여부를 판단한다[1]. 또한 빅데이터 분석을 통한 가짜 리뷰 필터링 시스템은 블로그 키워드, 검색 수, 신뢰도, 만족도 등을 분석해 정보를 제공하고, 블로그 게시물 수와 광고 게시물 수를 비교해 신뢰도를 평가한다[2]. 그러나, 이러한 시스템들은 주로 텍스트 데이터에 의존해 이미지 내 광고성 문구를 판별하지 못하는 한계가 있다.

본 논문에서는 블로그 텍스트 데이터뿐만 아니라 이미지에서도 OCR을 사용해 광고성 문구를 추출해 광고성 여부를 판단하는 시스템을 제안한다. 더불어 KoBERT 모델을 적용해 광고성을 더욱 정확히 식별하고, 이를 웹 브라우저 확장 프로그램 형태로 구현하였다.

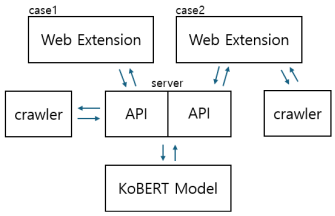

Fig. 1. System Structure

III. The Proposed Scheme

3-1. KoBERT 모델 기반 광고성 여부 판별 시스템

본 논문은 Fig 1과 같이 KoBERT 모델을 기반으로 네이버 블로그 글의 광고성 여부를 분류하는 웹 브라우저 확장 프로그램 시스템을 제안한다.

제안하는 시스템은 두 가지 방식으로 작동한다. 첫 번째 방식은 Web Extension이 서버의 API에 블로그 url을 포함한 요청을 보내고, 서버는 해당 블로그 url에 접속해 crawler를 통해 데이터를 수집한 후, KoBERT 모델로 광고성 여부를 판별한다. 두 번째 방식은 Web Extension에서 crawler를 통해 직접 블로그 페이지의 데이터를 수집하고, 이를 서버의 API에 전송해 KoBERT 모델로 광고성 여부를 판별한다.

각 Web Extension은 API와 통신하여 요청 및 응답을 주고받으며, 서버는 KoBERT 모델과 상호 작용해 최종 결과를 반환한다.

3-2. KoBERT 모델 학습을 위한 데이터 수집 및 학습

모델 학습에 필요한 데이터는 블로그 본문, 제목, 광고성 문구가 포함된 이미지를 크롤링하여 OCR을 사용해 텍스트를 추출하였다. OCR을 사용하는 이유는 네이버 블로그에서 광고임을 명시하는 이미지가 많이 사용되기 때문이다. "제공받았", "지원받았", "소정의", "파트너스" 등의 단어가 포함된 경우 1로, 그렇지 않은 경우 0으로 라벨링하여 총 6473개의 데이터셋을 확보하였다. 수집된 데이터셋을 전처리한 후, 한국어 특화 BERT 모델인 KoBERT를 활용해 학습시켰다. 학습된 모델은 웹 브라우저 확장 프로그램에 맞춰 파인튜닝 되었다.

3-3. KoBERT 기반 광고성 여부 판별 시스템 구현

본 논문에서 제안된 기법을 활용하여 광고성 여부 판별 시스템을 구현하였다. 사용자가 네이버 블로그 목록에서 링크를 우클릭하여 홍보성 체크 옵션을 선택하면 서버가 블로그 내용을 수집해 KoBERT 모델로 분류 후 결과를 반환한다(Fig 2). 또한 사용자가 블로그 페이지에서 확장 프로그램 아이콘에 있는 버튼을 클릭하면 현재 페이지 데이터를 서버에 전송해 KoBERT 모델로 분석한 결과를 표시한다(Fig 3).

Fig. 2. Example of the result screen of Method 1

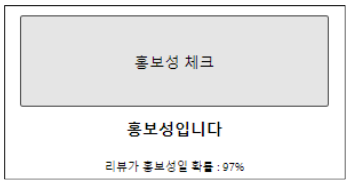

Fig. 3. Example of the result screen of Method 2

IV. Conclusions

제안된 시스템을 통해 블로그 글의 광고성 여부를 쉽게 확인할 수 있으며, 이를 통해 투명한 정보 제공과 공정한 브라우징 환경을 조성할 수 있다. 향후 연구에서는 시스템의 실시간 처리 속도를 개선하고, 사용자 인터페이스를 최적화하는 데 초점을 맞출 수 있을 것이다. 이를 통해 더 많은 사용자들에게 유용한 도구로 활용될 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

“본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 지원을 받아 수행되었음”(2019-0-01817)

REFERENCES

[1] Donggwan Yoo, Hangil Lim, & Dong-Kyu Chae, “A Fake Review Detection and Important Word Identification based on LSTM-Attention Models.” Academic Presentation Papers of the Korean Information Science Association, Jeju. pp.1,505- 1,506 June 2021.

[2] Jeong, Davichi, & Rho, Young-J, “Development of Filtering System ADDAVICHI for Fake Reviews using Big Data Analysis.” Journal of the Korea Internet Broadcasting and Communications Association , pp.1-8 June 2019.