Annual Spring Conference of KIPS

XAI를 활용한 실시간 허위광고 분류 서비스 개발

오수빈[1], 김주현[1], 곽소정[1], 조민수[1]
[1]광운대학교 정보융합학부

Introduction

❑ **현 허위 과대 광고 분류에 대한 한계점**

- 허위 과대 광고 관련 직접 법령에 따라 식약처에서 직접 분류
- 이에 따라, 허위 과대 광고에 따라 과도한 인력, 시간, 비용 요구

❑ **AI 기반 허위 과대 광고 분류 시스템의 필요성**

- 복잡, 애매한 법률 규정으로 인해 소비자, 기업에 직접 피해
- 새로운 형태의 허위 과대 광고 발생에 따른 적응성 부족

❑ **본 연구 목적**

- LSTM, XAI 기술을 활용한 실시간, 능동적 허위 과대 광고 위험도 분석 알고리즘 및 서비스 개발

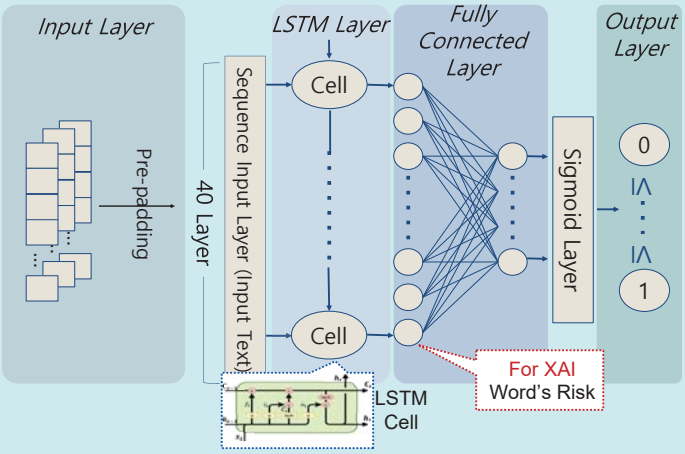
Method

❑ **데이터 처리 (Data Preprocessing and Word Embedding)**

- Label Data : 광고 글을 허위(0)/허용(1)으로 Labeling
- 토큰의 분포와 Label 기반 가중치로 워드 임베딩하여 XAI 적용

❑ **모델 아키텍쳐 (LSTM-based NN for XAI)**

- (Input) LSTM에 적합한 Pre-padding(size 40) 활용
- (LSTM) LSTM Layer를 train 문장마다 적용하여 모델 학습
- (FC) fully connected layer를 통해 시그모이드 활성함수에 전달
- (Output) Output이 0에 가까우면 허위광고, 1에 가까우면 허용광고로 분류



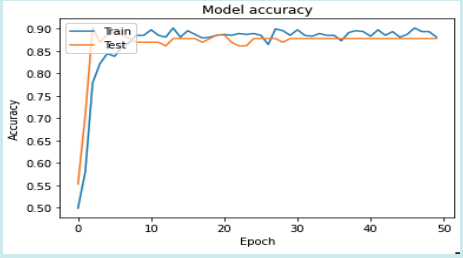
본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 지원사업의 연구결과로 수행되었음(2017-0-00096)

광운대학교 KwangWoon University

Results

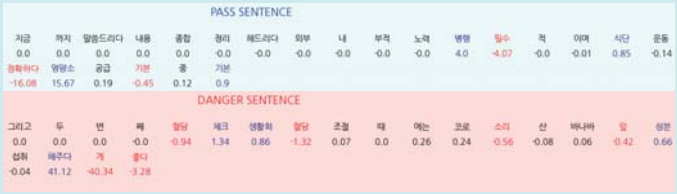
❑ **허위광고 분류 모델 정확성**

- Validation(split=0.2) test accuracy가 0.89로 높은 성능을 보임
- batch size=32로 train accuracy가 0.92, loss가 0.26로 학습됨



❑ **XAI 기반 원인 분석 결과**

- 허용 표현은 다음과 같이 파란색으로 표시되고, 허위 표현은 빨간색으로 표시되며, 각 단어들 아래 숫자 값은 문장 내 허위 허용의 영향도를 나타냄.



Conclusion

❑ 실시간 허위광고 위험 분석 서비스를 통해 수작업의 방식보다 **정확하고 빠른 허위광고 분류 가능**

❑ **XAI를 활용한 텍스트 내 허위광고 대상이 되는 원인을 제시**하여 소비자, 기업에게 유의미한 정보 제공이 가능

Reference

[1] Hyojeong Lee, Jiwan Lee, Youjeong Yang, Minji Cho, Bohyun Lee, Hyewon Lee, Yoonhee Kim.(2017).A Design of an Advertisement Analysis System using Deep Learning and Opinion Mining.한국정보과학회 학술발표논문집,2077-2079.

[2] Tae-Uk Yun, Hyunchul Ahn.(2018).Fake News Detection for Korean News Using Text Mining and Machine Learning Techniques.Journal of Information Technology Applications & Management,25(1),19-32.

[3] Yi, M. H., Lim, M. J., & Shin, J. H. (2020, September 30). Swearword Detection Method Considering Meaning of Words and Sentences. Korean Institute of Smart Media. Korean Institute of Smart Media.

ASK 2022

사단법인 한국정보처리학회 KIPS Korea Information Processing Society