Journal of KIIT. Vol. 21, No. 7, pp. 7-12, Jul. 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.7.7

자연어 처리 기법을 활용한 판례 기반 모욕성 문장 분석 시스템

김가영*, 공현정**, 유석종***

Legal Case-based Insult Sentence Analysis System using Natural Language Processing Techniques

Ga-Young Kim*, Hyeon-Jeong Kong**, and Seok-Jong Yu***

요 약

SNS가 발전함에 따라 온라인에서 타인을 공공연하게 모욕하는 경우가 증가하며 사회적 문제로 대두되고 있다. 포털사이트에서는 이를 막기 위해 데이터베이스 기반 비속어 필터링 방식으로 악성 댓글 등을 차단하고 있다. 하지만 맥락상 비속어가 아닌 단어를 욕설로 판단하거나 타인을 모욕하는 경우는 제대로 차단하지 못하고 있다. 본 연구에서는 자연어 처리 기법을 활용한 판례 기반 모욕성 문장 분석 시스템을 제안한다. 구현 시스템은 딥러닝 모델을 사용하여 판례 데이터셋을 학습하여 대상 문장의 모욕죄 확률을 추론하고, 텍스트 유사도 분석을 통해 근거 판례를 제시한다. 딥러닝 모델의 학습 데이터셋과 시험 데이터셋의 정확도를 고려하여 최종 모델을 선정하였으며, 웹 서비스를 구현하여 시스템 접근성을 높였다.

Abstract

With the advent of social media, there has been a significant rise in the number of individuals openly engaging in online insults, thereby emerging as a notable social issue. In order to protect users from unpleasant experiences, online portal sites employ database-driven profanity filters to render offensive content invisible. However, these filters often fall short in preventing instances where individuals insult others by judging certain words, not classified as profanity, as offensive within a specific context. Therefore, this paper aims to propose a precedent-based insult sentence analysis system, utilizing advanced natural language processing techniques. The system leverages a deep learning model, rooted in precedents, to infer the likelihood of guilt associated with an insult sentence. Furthermore, the system presents users with comparable precedents through similarity analysis. The ultimate model was chosen based on the accuracy of the test dataset and the training dataset while ensuring enhanced accessibility by deploying it on the web.

Keywords

insult case, classification, machine learning, BERT model, text similarity

* 숙명여자대학교 통계학과 학사과정
- ORCID: https://orcid.org/0009-0000-1903-0443
** 숙명여자대학교 소프트웨어학부 학사과정
- ORCID: https://orcid.org/0009-0007-3691-2350
*** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: May 21, 2023, Revised: Jun. 19, 2023, Accepted: Jun. 22, 2023

· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Womens's University,
Cheongpa-ro 100, 47-gil, Cheongpa-ro, Yongsan-gu, Seoul, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

Ⅰ. 서 론

SNS가 발전함에 따라 온라인 상에 업로드되는 글의 양이 기하급수적으로 늘어나고 있다. 포털사이트에는 매달 수십 건의 모욕죄 상담 사례가 올라오고 있으며 관련 범죄 역시 증가하고 있는 상황이다[1]. 모욕죄란 공연하게 사람을 모욕함으로써 성립하는 범죄로, 인터넷상에서 게시한 모욕성 문장은 마땅히 공연성을 인정받아 모욕죄가 성립할 가능성이 크다. 현재 이에 대한 방지책으로 대부분의 포털 사이트에서는 욕설 및 비속어 데이터베이스를 사용하여 비속어를 단어 매칭 방식으로 필터링하고 있다. 하지만 이는 욕설 단어를 가리기만 할 뿐 모욕성 문장의 재확산을 막지 못하며 특정 단어가 맥락상 비속어 또는 모욕성으로 쓰이지 않았더라도 이를 구분하지 못하는 단점이 있다. 이러한 문제점을 개선하기 위해서는 맥락에 따라 모욕성 문장을 분류하고, 모욕죄가 성립되는 문장들의 재생산을 막을 수 있도록 작성자에게 경고할 수 있는 시스템에 대한 연구가 필요하다. 따라서 본 연구에서는 기계학습 기반의 자연어 처리를 활용하여 모욕죄 판례들을 기반으로 모욕성 문장을 분류하고, 텍스트 유사도로 관련 판례를 확인할 수 있는 시스템을 구현하고자 한다.

본 연구에서는 법률 검색 서비스인 리걸서치(LegalSearch)[2]와 케이스노트(CaseNote)[3]에서 모욕죄 관련 판례를 수집한 후 자연어 처리 모델을 개발하여 모욕성 문장의 분류와 유사 판례 탐색 서비스를 구현하고자 한다. 기존 연구 중 판례 기반 말뭉치를 학습하는 머신러닝 모델에 관한 연구가 미흡하여 모욕죄 판례를 사용해 훈련 데이터셋을 구축하고 이를 바탕으로 다양한 분류 모델을 학습시켜 성능 비교 실험을 진행한다. 실험 모델로 로지스틱 회귀(Logistic regression), LSTM, Transformer 세 가지를 선택하고 분류 정확도를 비교한다. 또한 코사인 유사도를 사용하여 대상 문장과 유사한 판례를 탐색하는 기능을 구현한다.

본 논문의 2장에서는 자연어 처리 모델인 Transformer와 형태소 분석에 사용된 KoNLPy[4]와 KoBERT[5]를 소개한다. 3장에서는 제안하는 판례 기반 모욕죄 예측 시스템에 대해 기술하였다. 4장에서는 실험 수행 결과를 제시하고, 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

2.1 딥러닝 기반 비속어 방지 시스템

기존 관련 연구[6][7]는 주로 딥러닝 모델을 통해 인터넷 채팅과 같은 상황에서 악성 단어들을 필터링하는데 초점이 맞추어져 있다. 이 방식은 여전히 필터링 된 단어를 유추할 수 있는 경우가 많아 사용자들에게 경각심을 주기에 미흡하며 어떠한 이유로 딥러닝 모델이 결과를 도출했는지 사용자들이 알기 어렵다. 그러나 제안 시스템은 모욕죄 판례를 기반으로 학습하며 입력된 문장과 유사한 판례를 제공하기 때문에 사용자가 해당 결과를 쉽게 이해할 수 있도록 설계하였다. 따라서 기존 딥러닝 기반 비속어 방지 시스템의 단점을 극복할 수 있을 것으로 기대한다.

2.2 자연어 처리 기법

자연어처리 기법 중 본 연구에서 활용한 LSTM과 Transformer 모델에 대해 소개한다. 장단기 메모리(LSTM, Long Short-Term Memory)은 순환 신경망(RNN, Recurrent Neural Network) 기법의 하나로 셀, 입력 게이트, 출력 게이트, 망각 게이트를 이용해 기존 RNN의 문제인 기울기 소멸 문제(Vanishing gradient problem)를 방지한 모델로, 기존 RNN에 비하여 형태소 기반 텍스트 데이터에서 높은 정확도와 빠른 예측 결과를 제공한다[4][5].

BERT[8]는 2018년 Google의 Devlin이 제안한 전이 학습 방법으로, 사전 학습된 대용량의 레이블링 되지 않는 데이터를 이용하여 언어 모델을 학습하고, 이를 토대로 자연어 처리 작업을 위한 신경망을 추가하는 구조를 지니고 있다. 이러한 BERT 모델은 대용량 데이터를 통해 사전 학습할 경우 상대적으로 적은 자원만으로 높은 성능을 보인다. 그러나 영어 기반으로 사전 학습되어 있어 한국어의 경우 성능이 하락한다.

본 연구에서는 한국어 기반으로 사전학습된 SKTBrain의 KoBERT[4]를 사용하였다. 형태소 분석기로는 성능과 속도 모두 뛰어난 오픈소스 라이브러리인 KoNLPy[4]의 Okt를 사용하였다.

2.3 Transformer

Transformer[9]은 기존 seq2seq 모델의 정보 손실 문제를 개선한 모델로, 어텐션을 이용한 인코더 디코더의 구조로 이루어져 있다. 번역 성능에서도 RNN보다 우수한 성능을 보여 실험 모델로 선정하였다. Transformer의 디코더 구조로 설계된 GPT 모델도 고려하였으나 아래의 이유로 채택하지 않았다. 첫째, 비지도 학습으로 문장을 생성하는 모델이기 때문에 사용자가 신뢰할 수 있는 정보를 주기 어려우며, 둘째, 모욕성 문장을 학습한 GPT 모델이 이를 다른 질문의 답변으로 활용할 위험성이 있다.

III. 판례 기반 모욕성 문장 분석 시스템

3.1 시스템 구조

그림 1은 본 논문에서 제안하는 시스템의 전체 구조와 처리과정이다.

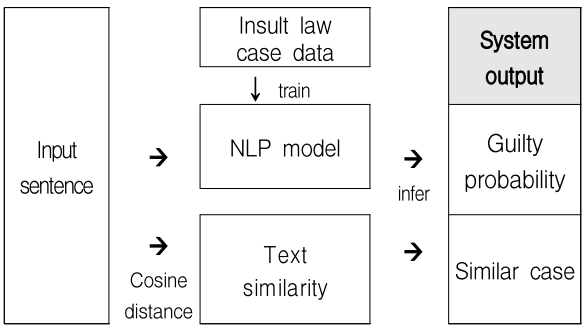

그림 1. 모욕성 문장 분석 시스템 구조
Fig. 1. Architecture of insult sentence analysis system

1) 데이터 크롤링을 통해 사전에 수집한 판례 데이터를 전처리한 후 데이터셋을 구축하여 모델 학습에 사용한다.
2) 사용자 문장을 입력받아 머신러닝 모델을 통해 모욕죄 유죄 확률을 예측한다.
3) 코사인 거리를 이용하여 유사 모욕죄 판례를 탐색한다.

3.2 데이터 수집 및 데이터셋 구축

모욕죄 관련 판례를 수집하기 위해 법률 검색 시스템인 리걸서치와 케이스노트 사이트를 선정하였다. Python Selenium 라이브러리를 사용한 동적 웹 크롤링을 통해 두 사이트로부터 사건명에 '모욕'이 포함된 판례들의 사건번호와 주문, 판례 전문을 추출했다. 이후 사실심 판결 등 범죄 사실이 드러난 판례만 저장하여 쟁점이 된 모욕 문장들을 추가로 파악하였다. 각 판례의 주문을 확인하여 유무죄 여부를 구분했다. "항소를 기각한다", "원심판결을 파기한다" 등 주문에 구체적인 형벌 사항이 나타나지 않은 판례는 이전 판결을 확인하여 유무죄 여부를 수기로 저장하였다. 중복 판례 제거 후 187개의 판례 데이터를 우선 수집하였다. 수집한 판례 데이터는 총 403개의 문장, 3827개의 어절로 이루어져 있다. 이중 유죄 판례는 152개, 무죄 판례는 35개이다.

또한 머신러닝 모델의 학습 성능을 높이기 위해 다음의 3가지 텍스트 증강 작업을 수행하였다[10]. 먼저, Random Swap은 특정 단어들의 위치를 바꾸는 기법으로 단어의 위치에 따라 모델이 편향적으로 학습하는 것을 방지하기 위해 진행하였다. 두 번째로 Translation은 문장을 특정 언어로 번역한 후 다시 역번역하는 증강 기법이다. 데이터셋에 사용된 유사한 단어도 모델 학습에 사용하여 일반화 성능을 높이기 위해 사용하였다. Random Insertion은 단어 사이에 특정 단어를 삽입하는 증강 기법으로, 모델이 여러 파생 문장을 학습할 수 있도록 한다. 구현 방법으로는 앞에서 설명한 텍스트 증강 기법을 한국어에 맞춰 구현한 ktextaug 패키지[11]를 활용하여 진행하였다. 그 결과 187개이던 판례 데이터를 935개로 증강하는데 성공하였다. 하지만 유죄 데이터 760개, 무죄 데이터 175개로 클래스 불균형이 여전히 존재하였고, 이를 해결하기 위해 AI Hub[12]의 한국어 SNS 데이터셋을 샘플링하여 무죄 데이터를 추가하여, 최종적으로 1,520개의 실험 데이터셋을 구축하였다.

표 1. 판례 데이터
Table 1. Information of insult law case data set

	Augmentation technique	Count
Guilt	Random swap * 2 translation random insertion	760
Innocence		760

데이터셋의 전처리는 다음과 같이 진행하였다. 첫째로 정규 표현식을 수행하여 중복되는 문장을 제거하고 공백 값을 제거하였다. 둘째로 단어를 토큰화하고 형태소를 분석하였다. 셋째로 불용어 사전을 이용하여 불용어를 정의하고 해당하는 토큰을 처리하였다.

3.3 실험 모델

본 시스템은 판례 데이터를 기반으로 학습된 모델을 통해 대상 문장의 모욕죄 성립 가능성을 예측하는 기능을 제공한다. 이를 위해 세 가지 분류 모델을 선정하여 모욕죄 유죄 혹은 무죄로 라벨링 된 데이터셋을 학습시켜 정확도를 비교하였다. 실험을 위해 3.2절에서 구축한 판례 데이터셋을 훈련, 검증, 시험 세트(5:3:2)로 분할하여 사용하였다.

로지스틱 회귀는 모욕죄 예측을 위한 기준 실험 모델로 선정하였고, 사이킷런의 선형모델 라이브러리를 사용하여 통해 구현하였다. L2 규제를 활용하여 과적합을 방지하였으며, C가 1.0일 때, 시험 데이터셋의 성능이 제일 높아 해당 파라미터를 최적으로 설정하였다. LSTM 모델은 RNN 모델보다 형태소 기반 텍스트 데이터에서 높은 성능을 보이기 때문에 선택하였다. 임베딩층을 100개로 두었으며, 은닉층을 128개로 두었다. 출력층에는 시그모이드 활성화 함수를 사용하였다. 에포크는 50으로 설정하였으며, 0.5724에서 시작한 훈련 데이터의 손실은 0.0356일 때 수렴하는 양상을 보였다. Transformer 모델은 한국어 텍스트 데이터셋에 대하여 가장 높은 성능을 보이는 모델로, 768개의 은닉층을 두었으며, Dropout 기법을 0.1로 적용하여 과적합을 방지하였다. 배치 크기는 64이며 학습률은 5e-5이다. 일반적으로 높은 성능을 보이는 옵티마이저 AdamW와 CrossEntropyLoss를 활용하였다. 그레디언트 클리핑을 사용하여 1로 두고 학습률을 최적화하였다. 에포크는 50으로 설정하였으며, 에포크 43부터 학습 데이터의 손실이 0.0014로 수렴하였다.

3.4 유사 판례 탐색

본 시스템은 사용자로부터 입력된 대상 문장과 비슷한 맥락으로 모욕죄 판결을 받은 판례를 탐색할 수 있다. 판례 데이터셋 내의 모욕 문장들과 대상 문장 간의 코사인 유사도를 구하여 유사 판례를 탐색하였다. KoNLPy의 Okt를 사용해 모든 문장을 토큰화하고, scikit-learn의 TfidfVectorizer를 사용해 입력된 문장과 판례 데이터셋 벡터들의 코사인 유사도를 계산하여 유사 판례를 10개를 추출한다.

Ⅳ. 시스템 구현 및 성능평가

4.1 성능평가 실험 결과

표 2와 같이 훈련 데이터셋에 대한 정확도는 Transformer와 로지스틱 회귀가 가장 높았다. 시험 데이터셋 정확도는 Transformer, LSTM, 로지스틱 회귀의 순서이다. 학습 데이터의 손실까지 고려하면 Transformer 모델이 0.0014로 가장 좋은 모델이라고 볼 수 있다. 그러나 Transformer 모델은 서비스로 구현했을 때 응답 속도가 1분 이상 길어지는 문제점이 있었다. 이중 LSTM 모델에 대해 실제 사례 대비 정확도를 확인하기 위한 정성평가를 진행하였다. 표 3은 네이버 지식in에 올라온 모욕죄 관련 질문 [13]-[15]과 실제 변호사의 답변을 자체 각색하여 실험한 결과이다. 두 예측 결과가 대체로 일치하는 것을 확인할 수 있다.

표 2. 예측 정확도 비교
Table 2. Comparison of prediction accuracy

Model	Train accuracy	Test accuracy	Train loss
Transformer	99.6%	97.6%	0.0014
LSTM	97.8%	93.8%	0.0356
Logistic regression	99.7%	92.7%	0.5928

표 3. 변호사와 실험 모델의 모욕죄 예측 비교
Table 3. Insult prediction between lawyers and model

Target sentence	Lawyer expectation	System prediction
너 이새* 망상장애있는 스토커 새*아니냐	High	99.1%
벌레랑 겜하기 싫다 벌레* 손가락 ***남	Low	19.0%
진짜 운전 너무 못하시네 난 못하겠으니까 니가 가르쳐	Low	6.5%

모욕죄의 성립 요건에는 모욕의 대상이 누구인지 인지할 수 있어야 한다는 '특정성'이 존재한다. 표 4와 같이 맥락이 비슷하지만, 특정성의 정도가 다른 문장들을 실험하였고, 특정성이 짙은 문장일수록 모욕죄 유죄 확률이 높게 나타나는 것을 알 수 있다.

표 4. 실험 모델의 특정성 반영 검증
Table 4. Validation of specificity in model

Target sentence	Specificity	Prediction
이 글을 쓴 사람은 사기꾼입니다. 제게 돈을 빌리고 갚지 않았습니다. 엄청난 악질 사기꾼입니다.	High	74.1%
야 이 사기꾼 같은 놈아	Medium	37.7%
사기꾼	Low	1.7%

4.2 유사 판례 탐색

표 5는 유사 판례를 탐색한 결과이다. 본 연구 결과를 웹에서 활용할 수 있도록 웹사이트를 개발하였다. 웹 서버는 Python 기반 웹프레임워크인 Flask를 사용하였고, MariaDB 데이터베이스를 사용하여 판례 데이터셋을 저장하였다.

표 5. 유사 판례 탐색 결과
Table 5. Result of text similarity

Target sentence	Similar law case	Target sentence
기레기님 소설 쓰네... 가짜뉴스 퍼트리지 마세요	대구지방법원 2021.7.20. 선고 2021노 1035 판결[모욕]	이런 걸 기레기라고 하죠?
이 글을 쓴 사람은 사기꾼입니다. 제게 돈을 빌리고 갚지 않았습니다. 엄청난 악질 사기꾼입니다.	제주지방법원 2017.4.27. 선고 2016노544 판결 [명예훼손, 신용훼손, 업무방해, 모욕]	사기를 쳐 회사를 인수하였다. 사기꾼

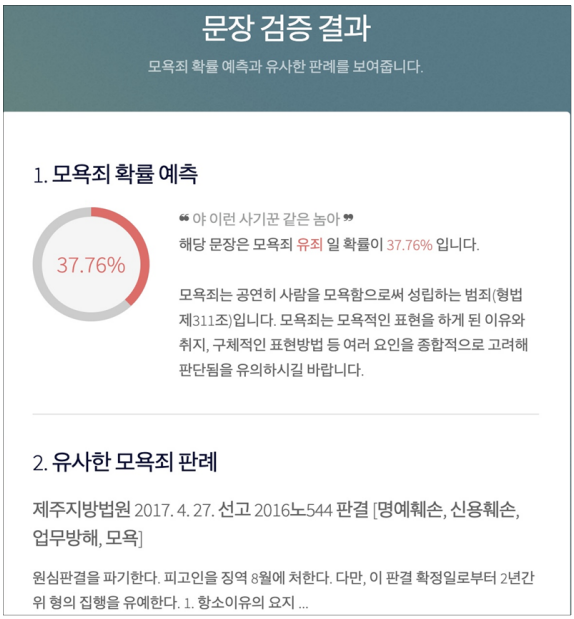

그림 2. 유사 판례 검색 결과
Fig. 2. Search result of similar case sentence

V. 결론 및 향후 과제

본 연구에서는 모욕성 예측에서 기계 학습의 효과를 확인하고자, 판례 데이터셋을 구축하여 모욕죄 예측 및 유사 판례 탐색 시스템을 구현하였다. 부족한 판례 기반 데이터를 증강하여 로지스틱회귀, LSTM, Transformer 모델을 사용하여 예측 성능 평가 실험을 수행하였다. 본 연구의 한계점으로 과거 판례를 모델 학습에 사용하기 때문에 유사 판례가 없는 경우 콜드 스타트(Cold start) 문제가 발생할 수 있다. 또한, 발화 문장의 모욕성만 파악할 뿐 공연성 등 사건의 전체적인 맥락 판단은 제공하지 않고 있다. 후속 연구에서 판례 데이터를 보강하거나 모욕성이 높은 단어들의 가중치 조정을 통해 개선을 기대할 수 있을 것이다.

References

[1] NAVER Legal Consultation, https://cafe.naver.com/hacknate [accessed: May 4, 2023]
[2] LegalSearch, https://legalsearch.kr [accessed: May 4, 2023]
[3] CaseNote, https://casenote.kr [accessed: May 4, 2023]

[4] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python", Proc. of the 26th Annual Conference on Human & Cognitive Language Technology, pp. 133-136, Oct. 2014.

[5] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, "KR-BERT: A Small-Scale Korean-Specific Language Model", arXiv:2008.03979, Aug. 2020. https://doi.org/10.48550/arXiv.2008.03979.

[6] J. H. Lee, D. G. Han, H. Y. Kim, and B. K. An, "An Internet Vulgarity Filtering System based on Machine Learning", ICEIC(2018), Vol. 41, No. 2, pp. 852-853, Nov. 2018.

[7] Y. Ha, J. Cheon, I. Wang, M. Park, and G. Woo, "A Filtering Method of Malicious Comments Through Morpheme Analysis", Vol. 21, No. 9, pp. 750-761, Sep. 2021. https://doi.org/10.5392/JKCA.2021.21.09.750.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805, pp. 3-5, May 2019. https://doi.org/10.48550/arXiv.1810.04805.

[9] A. Vaswani, et al, "Attention Is All You Need", arXiv:1706.03762, Dec 2017. https://doi.org/10.48550/arXiv.1706.03762.

[10] J. Wei and K. Zou "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", arXiv:1901.11196, Aug. 2019. https://doi.org/10.48550/arXiv.1901.11196.

[11] Ktextaug, https://github.com/jucho2725/ktextaug [accessed: May 4, 2023]

[12] AIhub Korean SNS Dataset, https://aihub.or.kr [accessed: May 4, 2023]

[13] Naver case example 1, https://kin.naver.com/qna/detail.naver?d1id=6&dirId=60215&docId=438514354 [accessed: May 4, 2023]

[14] Naver case example 2, https://kin.naver.com/qna/detail.naver?d1id=6&dirId=60206&docId=423979997 [accessed: May 4, 2023]

[15] Naver case example 3, https://kin.naver.com/qna/detail.naver?d1id=6&dirId=60206&docId=401154420 [accessed: May 4, 2023]

저자소개

김 가 영 (Ga-Young Kim)

2020년 2월 ~ 현재 :
숙명여자대학교 통계학과
학사과정
관심분야 : 생성모델, 자연어처리

공 현 정 (Hyeon-Jeong Kong)

2018년 2월 ~ 현재 :
숙명여자대학교 소프트웨어학부
학사과정
관심분야 : 프로그래밍, 자연어처리

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교
컴퓨터과학과(이학사)
1996년 2월 : 연세대학교
컴퓨터과학과(이학석사)
2001년 2월 : 연세대학교
컴퓨터과학과(공학박사)
2005년 ~ 현재 : 숙명여자대학교
소프트웨어학부 교수
관심분야 : 데이터마이닝, 추천시스템, 정보시각화