Deep Learning을 이용한 악의적 댓글 탐지 모델들의 성능 비교

Comparative Study of Malicious Comment-detection Models based on Deep Learning

저자
(Authors)

안시후, 황석형, 김응희, 김민경
Sihu Ahn, Suk-Hyuong Hwang, Eung-Hee Kim, Minkyoung Kim

출처
(Source)

한국정보과학회 학술발표논문집 , 2019.12, 1478-1480 (3 pages)

발행처
(Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09301972

APA Style

안시후, 황석형, 김응희, 김민경 (2019). Deep Learning을 이용한 악의적 댓글 탐지 모델들의 성능 비교. 한국정보과학회 학술발표논문집, 1478-1480.

Deep Learning을 이용한 악의적 댓글 탐지 모델들의 성능 비교*

안시후○, 황석형, 김응희, 김민경+
선문대학교, 글로벌소프트웨어학과
{lmgtt, shwang, ehkim, minkyoungkim}@sunmoon.ac.kr

Comparative Study of Malicious Comment-detection Models based on Deep Learning

Sihu Ahn○, Suk-Hyuong Hwang, Eung-Hee Kim, Minkyoung Kim+
Department of Global Software Engineering, Sunmoon University

요 약

최근 악의적 댓글로 인한 연예인 자살 문제가 사회적으로 대두됨에 따라, 단순 욕설 필터링보다는 주관성이 배제될 수 없는 악성 댓글 판단 연구에 관심이 집중되고 있다. 이에, 본 논문에서는 네이버 뉴스 기사에서 수집된 댓글 데이터를 [모든 형태소 자소분리], [모든 형태소], [명사 자소분리], [명사] 네 가지 처리 방식으로 분류하여 Deep Learning 기법 중 자연어처리에 전통적으로 쓰인 RNN(Recurrent Neural Networks)과 LSTM(Long Short Term Memory) 그리고 비교적 최근 자연어처리에 효과적인 CNN(Convolutional Neural Network)을 악성 댓글 탐지에 적용하여 그 성능을 비교, 분석하고자 한다.

1. 서 론

최근 악의적 댓글로 인한 연예인 자살 문제가 사회적으로 대두되고 있다. 악의적 댓글의 영향으로 피해자는 정신적 피해를 받고 심한 경우 자살로 이어진다. 이에 따라, 가해자는 법적 처벌을 받게 되고, 악의적 댓글이 실린 사이트 또한 피해자에게 손해배상을 해야 한다. 이러한 피해 사례들과 법적으로 처벌이 되는 규정이 있으나 악의적 댓글을 포함한 사이버 명예훼손 피해는 그림 1에서 보이는 바와 같이 5년 새 2배나 증가하였다.

이러한 문제들로 인해 인터넷 실명제, 댓글 차단, 악의적 댓글 법 강화 등 다양한 의견이 재조명되고 있다. 기업들의 악의적 댓글 대처 현황들을 살펴보면,

그림1. 2014~2018 사이버 명예훼손(출처:대검찰청)

네이버는 클린봇을 통해 욕설을 필터링해주는 서비스를, 카카오는 댓글 차단, 유튜브는 댓글 삭제를 시행하고 있다. 하지만, 욕설이 직접적으로 드러나지 않더라도 비난성 댓글은 주관적 견해가 포함된 악의적 댓글이기 때문에 단순 욕설보다 판단하기 어렵다는 문제가 있다. 또한, 욕설 탐지 연구에 비해 악의적 댓글 탐지 연구가 활발하게 이루어지지 않고 있는 실정이며, 정확도 또한 상대적으로 낮은 성능을 보여주고 있다.

이에 따라, 본 논문에서는 악의적 댓글 탐지 정확도 향상을 위한 최근의 딥러닝 방법을 비교 분석한다. 그 대상으로 욕설 분류에 뛰어난 성능을 보이는 (1)CNN, 악의적 댓글 관련 선행연구에서 사용되고 있는 (2)RNN과 (3)LSTM 모델들을 악의적 댓글 탐지 모델에 적용하여 비교, 분석한다.

본 논문의 분석 결과는, 향후 악의적 댓글 차단 연구에 유용하게 활용될 수 있을 것으로 기대된다.

2. 관련 연구

김묘실, 강승식이 SVM을 이용한 악성 댓글 판별 시스템을 제안하였다[1]. 이 논문에서는 이진 분류기인 SVM을 통해 9:1의 데이터 셋으로 실험한 결과 68%의 정확도를 얻었다.

김진우 외 2명이 인공신경망을 적용한 악성 댓글 분류 모델들의 성능을 비교하였다[2,3,4]. 이 논문에서는 네이버 뉴스 10개 기사의 13,362개의 댓글을 활용하여 네 가지 품사 처리 방식([명사], [명사+형용사], [명사+형용사+동사], [모든 품사])과 RNN, LSTM, GRU 세 가지 알고리즘을 적용해 12개의 모델의 성능을 비교하였다.

* "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음"(2018-0-01865)
○ 발표자, + 교신저자(Email: minkyoungkim@sunmoon.ac.kr)

그 결과 전체 데이터 셋 기준 최대 96.6%의 성능, 1:1 데이터 셋 기준 최대 80.2%의 정확도를 보였지만 전체 데이터가 약 98:2로 악의적 댓글이 약 2%인 상태에서 정확도를 기준으로 작성되어 신뢰하기 어렵다. 예를 들어, 모두 다 악의적 댓글이 아니라고 한다면 정확도가 약 98%로 측정되는 문제가 있다.

박성희 외 2명이 CNN을 사용하여 온라인 게임에서의 욕설 탐지 모델을 제안하였다[5,6,7,8,9]. 이 논문에서는 온라인 게임 'ArcheAge'의 채팅 데이터를 통해 자소 및 음절 단위 분리한 데이터 셋을 CNN에 각각 적용하여 87%의 정확도를 보였다.

선행 연구들을 통해서 이진 분류 모델들은 non-linear문제를 차원 왜곡을 통해 해결하지 못한다면 낮은 성능 결과를 나타내는데, 이는 Deep Learning 모델이 자동으로 해결해 줄 수 있어 악의적 댓글 탐지에 유리하다는 것을 알 수 있었다. 또한, 한글의 악의적 댓글 탐지에 CNN을 적용한 사례가 없어, 본 연구에서는 이를 적용하여, RNN과 LSTM 모델의 성능과 비교, 분석하고자 한다.

3. 제안 모델 및 실험

3.1 데이터 수집

데이터는 R로 작성된 네이버 뉴스 트롤링을 위한 도구 N2H4[10]를 활용하여 2019년 9월 1일부터 10월 30일 네이버 기사에서 설리 관련 기사에서 댓글 약 40,000만 개를 수집하였다.

3.2 전처리

정진수 외 2명의 "명예훼손죄·모욕죄에 대한 판례의 판단 기준 연구[11]"에 따라 약 40,000개의 댓글 데이터 중 공연성, 모욕성, 특정성을 만족하는 악의적 댓글 823개를 추출하였다. 이후 빈도수가 1인 단어를 제거한 후 600개의 악의적 댓글 데이터를 추출하였고 일반 댓글 600개를 랜덤으로 추출하여 총 1,200개의 데이터를 1:1 비율로 생성하였다. 한국어 정보처리를 위한 파이썬 패키지인 KoNLPy[12]와 한글 자모분리/조합 작업을 위한 툴킷인 hangul-toolkit을 이용하여 명사, 모든 품사 두 가지 형태소로 분류하고 자모분리를 하여 네 가지 데이터 셋을 생성하였다. Keras의 Toknizer를 이용하여 데이터를 토큰화 및 벡터화하고 단어 벡터의 최대값으로 나누어 스케일링 후 단어의 최대 길이로 패딩 작업을 해주었다. 80:20rule을 바탕으로 8:2로 실험 데이터를 구성하고 확인은 5-fold로 나누어 학습할 때 사용하고자 하였다. 데이터 별 가장 긴 댓글 수와 32차원으로 임베딩하여 데이터 전처리를 완료하였다.

3.3 Deep Learning 모델 구현 및 실험

RNN, LSTM, CNN을 네 가지 데이터에 적용하여 정확도와 1epoch의 평균 학습속도를 비교하였다.

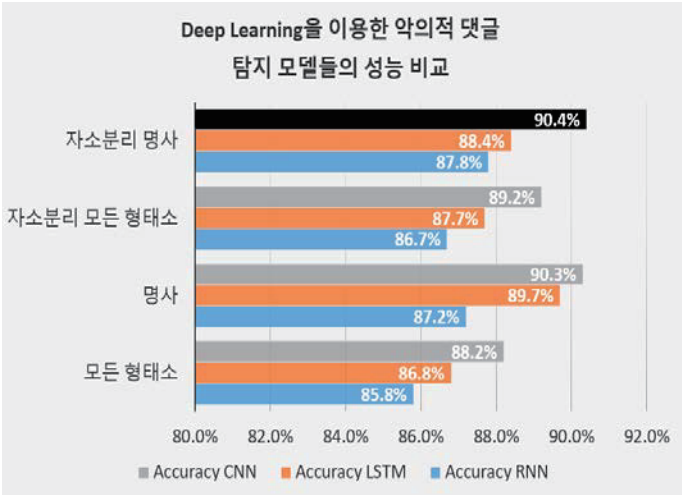

그림2. 본 연구 모델들의 정확도 비교

3.3.1 RNN 모델 구현

TensorFlow의 simpleRNN을 이용하여 구현하였으며 Hidden size는 32, 하나의 정답을 찾기 때문에 Dense의 Unit은 1, activation function은 sigmoid로 구성하였다.

3.3.2 LSTM 모델 구현

TensorFlow의 LSTM을 이용하여 구현하였으며 형태별 단어의 최대 길이와 activation function tanh로 구성하였고 Dense의 Unit은 1, activation function은 relu로 구성하였다.

3.3.3 CNN 모델 구현

TensorFlow의 Conv1D를 이용하여 구현하였으며 filter는 32, kernel size는 3, activation function은 relu로 구성하였고 Dense의 Unit은 1, activation function은 sigmoid로 구성하였다.

3.4 Compile & Fit

Loss는 binary cross entropy, optimizer는 adam을 이용하여 compile 하였으며, batch size는 60, epochs는 RNN:15, LSTM:7, CNN:17로 구성하여 네 가지 데이터별 fit 하여 모델 학습을 진행하였다.

4. 실험 결과

그림2가 보여주는 바와 같이, 모든 형태소 데이터는 필요 이상의 정보를 보유하여 자소분리 명사 데이터를 사용한 모델들이 뛰어난 성능을 보여주고 있으며, 그 중 CNN이 가장 우수한 성능을 나타낸다. 그림3은 정확도가 직전 선행연구(base line 2) 대비 10.21% 향상된 결과를 나타내고, 그림4는 학습 속도가 CNN이 RNN 대비 1.75배, LSTM 대비 14.2배 빨라짐을 보여준다.

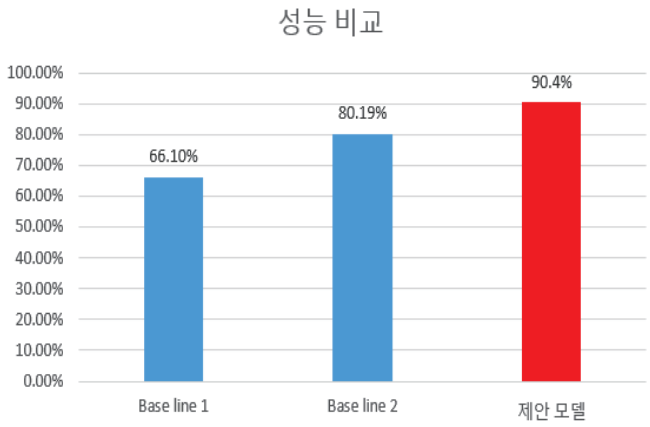

그림3. 선행연구들과의 정확도 비교

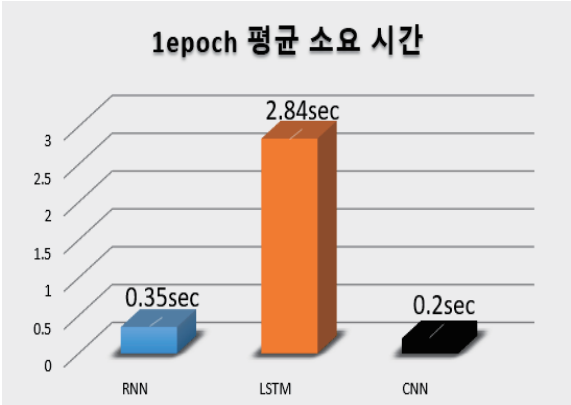

그림4. 학습 속도 비교

추가적으로, 데이터에 출현 빈도가 1회인 단어를 제거해줌으로써 모델의 정확도가 향상되는 결과를 얻을 수 있었다. 또한, 자소분리 시 사전의 크기를 기존보다 두 배 이상 줄일 수 있을 뿐만 아니라 유사도 측정할 수 있는 기회를 마련할 수 있어 향후 관리의 편의성 및 성능 향상에 기여할 수 있을 것이라고 생각된다.

5. 결론 및 향후 연구 계획

본 연구에서는 RNN, LSTM, CNN 3가지 Deep Learning 모델을 적용하여 모든 형태소, 모든 형태소 자소분리, 명사, 명사 자소분리 네 가지 처리 방식을 거친 데이터를 통해 총 12가지 모델들의 성능을 비교하였다. 직설적인 욕설과 다르게, 암묵적인 악의적 댓글은 그 변형이 쉬워 필터링 하는 것에 어려움이 있다. 따라서, 일괄적인 댓글 차단보다는 Deep Learning을 이용한 악의적 댓글 탐지 모델을 적용하여 댓글을 필터링할 수 있는 서비스가 필요 할 것으로 생각된다.

향후 주요 연구 방향은, 악의적 댓글의 단순 이진 분류보다는, 악의성을 확률적으로 나타냄으로써 사용자가 그 기준을 결정할 수 있도록 자율성을 제공하고자 한다.

참고문헌

[1]김묘실, 강승식, “SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현”, 한국정보과학회 언어공학연구회 학술발표 논문집, p.285-289, 2006.10
[2]김진우, 조혜인, 이봉규, “인공신경망을 적용한 악성 댓글 분류 모델들의 성능 비교”, 한국디지털콘텐츠학회 논문지 20(7), p.1429-1437, 2019.7
[3]colah’s blog, “Understanding LSTM Netwroks”, http://colah.github.io/posts/2015-08-Understanding, 2015.8
[4]엉드루, “딥러닝하기 7편. LSTM을 이용한 뉴스 분류하기“, https://m.blog.naver.com/PostView.nhn?blogId=htk1019&logNo=221255254613&proxyReferer=https%3A%2F%2F” 2018.4
[5]박성희, 김휘강, 우지영, “딥러닝을 사용한 온라인 게임에서의 욕설 탐지”, 한국컴퓨터정보학회 하계학술대회 논문집 27(2), p.13-14, 2019.7
[6]colah’s blog, “Conv Nets: A Modular Perspective” http://colah.github.io/posts/2014-07-Conv-Nets-Modular/ 2014.7
[7]colah’s blog, “Understanding Convolutions” http://colah.github.io/posts/2014-07-Understanding- 2014.7
[8]Yoon Kim, “Convolutional Neural Networks for Sentence Classification”, arXiv:1408.5882v2 [cs.CL], 2014.9
[9]WILDML, “Implementing a CNN for Text Classification in TensorFlow”, http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/ 2015.12
[10]Chan-Yub Park, “네이버 뉴스 수집을 위한 도구”, https://forkonlp.github.io/N2H4/
[11]정진수, 강태경, 김형길, “명예훼손죄·모욕죄에 대한 판례의 판단 기준 연구: 최근 10 년간(2005~2015)의 판례를 중심으로”, 형사정책연구원, 연구총서, 2015.12
[12]박은정, 조성준, “NoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014