「정보시스템연구」 제29권 제3호
한국정보시스템학회
2020년 9월, pp. 103~117

http://dx.doi.org/10.5859/KAIS.2020.29.3.103

하이웨이 네트워크 기반 CNN 모델링 및 사전 외 어휘 처리 기술을 활용한 악성 댓글 분류 연구*

이현상** · 이희준*** · 오세환****

〈목 차〉

Ⅰ. 서 론

정보통신 기술의 발달로 최근 국민 인터넷 이용률이 2019년 기준 약 92%로 증가하는 추세이다(과학기술정보통신부, 2020). 이에 따라 명예 훼손, 인신 공격, 사생활 침해 등의 문제를 유발하는 악성 댓글이 최근 큰 사회적 문제로 나타나고 있다. 악성 댓글로 인한 사이버 명예 훼손 사건 발생 및 검거 건수는 각각 2019년 기준 16,633건, 11,632건으로 점점 증가하고 있다(경찰청, 2020). 실제로 우리나라의 온라인 뉴스 소비 비율은 다른 국가들 보다 높은 수준이다(Reuters Institute, 2020), 이에 따라 우리나라의 포털 사이트 인터넷 여론 형성 수준은 타 국가에 비해 상대적으로 높아 악성 댓글의 영향력이 더욱 문제가 되고 있다.

이러한 상황에서 정부는 인터넷 실명제를 도입하려 했으나 2012년 헌법재판소에서 정보통신망 이용촉진 및 정보보호 등에 관한 법률 제44조의5 제1항 제2호 등의 근거로 위헌 판결을 받았다(김현귀, 2013). 판결에 따르면 인터넷 실명제가 국민의 표현의 자유를 위축시키고 개

* 이 논문은 과학기술정보통신부가 주최하고 과학기술정보통신부의 정보통신진흥기금으로 정보통신산업진흥원이 지원하는 개방형 경진대회 플랫폼 구축 사업의 ‘2020년 인공지능 온라인 경진대회 우수 성과 기업 사업화’ 사업지원을 받아 수행된 결과임 [과제번호: A0712-20-1021]
** 경북대학교 경영학부, coolwin20@knu.ac.kr(주저자)
*** 계명대학교 경영정보학과, hjlee@stu.kmu.ac.kr
**** 경북대학교 경영학부, sehwano@knu.ac.kr(교신저자)

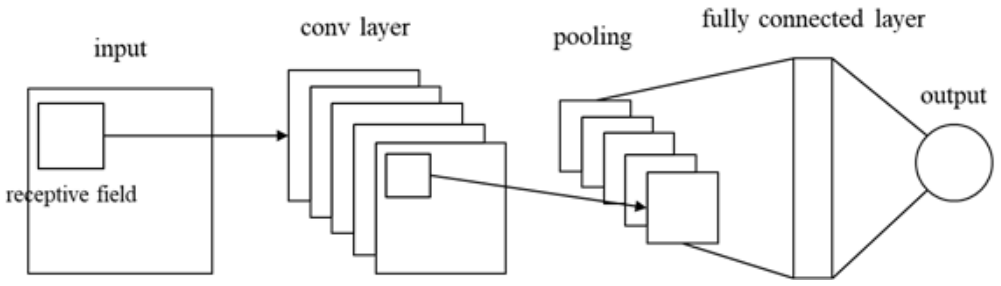
인정보자기결정권을 침해하는 등의 문제가 있다는 것이다. (주)한국리서치에서 발표한 자료에 의하면 인터넷 웹사이트 댓글을 읽는 비율이 86%, SNS에서 61%, 온라인 동영상 플랫폼에서 68%로 나타났다((주)한국리서치, 2019). 인터넷 사용자들이 댓글을 읽는 비율이 높은 만큼 일방적으로 댓글 기능을 차단할 경우 표현의 자유 침해, 사용자 만족도 저하 등의 문제를 초래할 수 있다. 인공지능 기술 기반으로 효과적인 악성 댓글 분류 모델링을 개발할 수 있다면 악성 댓글 문제를 상당부분 해결할 수 있을 것이다.

악성 댓글 문제를 해결하기 위해 국내외에서 다양한 머신러닝 기법을 기반으로 분류 모델링을 연구했다(배민영 등, 2009; 김현정 등, 2011; 홍진주 등, 2016; Georgakopoulos et al., 2018; Li, 2018; Srivastava, 2018; Moon et al., 2020). 최근 국내 경영정보 분야에서도 머신러닝 기법을 활용한 텍스트 마이닝, 예측 모델, 그리고 추천 시스템 등의 연구들이 다양한 방식으로 수행되고 있다(Park and Ha, 2017; 양낙영 등, 2018; 정건용 등, 2019). 국외 연구 사례에서는 딥러닝 기술 기반 악성 댓글 분류 모델링 연구가 어느 정도 수행되었으나 한국어 텍스트 대상 연구는 아직까지 부족한 상황이다(Georgakopoulos et al., 2018; Li, 2018; Srivastava, 2018). 기존 연구의 경우 단어 사전을 구축하는 방법으로 데이터가 부족하고 문맥적 상황을 고려하지 못하며 신조어에 대한 대처가 미흡하고 다양한 악성 댓글의 유형을 고려하지 못한다는 문제가 있다(배민영 등, 2009; 김현정 등, 2011; 홍진주 등, 2016; Georgakopoulos et al., 2018; Moon et al., 2020). 본 연구의 목적은 기존 연구의 문제점을 해결하기 위해 사전 외 어휘(Out of Vocabulary) 처리에 대한 접근 방식과 딥러닝 모델링 기법을 제시하여 분류 모델의 성능을 높이고, 악성 댓글의 세분화된 유형을 분류할 수 있는 모델을 개발하는 것이다. 이를 통해 건전한 인터넷 교류 문화에 기여할 수 있을 것으로 기대한다.

Ⅱ. 문헌 연구

2.1 CNN

악성 댓글을 효과적으로 분류하기 위해 국내외에서 CNN(Convolution Neural Network)기반 모델링을 활용하는 연구들이 우수한 성과를 달성했다(Georgakopoulos et al., 2018; Srivastava, 2018). 악성 댓글의 데이터 특성상 텍스트의 길이가 짧고 변칙적이며 분류 문제기 때문에 문장의 시퀀스(sequence)보다는 비선형적인 특성을 제대로 학습할 수 있는 CNN 분류 기법이 적합하다고 볼 수 있다. CNN이란 <그림 1>처럼 DNN(Deep Neural Network)의 1차원적 입력 배열의 한계를 극복하기 위해 이미지의 공간 정보를 유지한 채로 학습이 가능한 합성곱 계층(convolutional layer) 기반 신경망 학습 모델을 의미한다(Lecun et al., 1990). 합성공 신경망 분류 모델링에서는 입력 데이터를 수용 영역 (receptive field)으로 분할하여 합성곱 계층을 구축하고, 풀링(pooling)을 통해 픽셀(pixel) 값을 한정하여 차원 축소를 한 다음 완전 연결 계층(fully connected layer)을 통해 분류 값을 산출한다. 세부적으로 합성곱 계층에서는 기존 신경망 분류와는 다르게 입력 및 출



<그림 1> CNN 아키텍처

력 값을 3차원으로 유지하는 것으로 3차원 이미지 데이터에 대해서 전부 뉴런과 연결하는 것이 아니라 수용영역(receptive field)에 할당된 픽셀에만 연결한다. 일반적으로 입력 데이터에 사용되는 이미지 데이터는 가로, 세로, 채널(channel)에 해당하는 3차원 데이터가 사용된다. 초기 합성곱 계층에서는 수용영역의 픽셀에 대해서 학습하고 완전 연결 계층에서 전체적인 패턴을 인식하는 것이다. 수용영역은 CNN에서 필터(filter) 혹은 커널(kernel)이라고 명명되며 이미지 픽셀 데이터를 분할하여 합성곱 계층을 구성한다. 합성곱 계층 연산은 입력 데이터에서 수용영역의 윈도우(window)를 일정한 간격으로 이동하면서 계산한다. 입력 데이터와 수용영역 사이의 대응하는 값을 곱하여 집계하고 편향 값을 더한다. 합성곱 계층 연산 시에는 이미지 데이터의 모서리 부분 픽셀 데이터의 신호가 희미해질 수 있기 때문에 패딩(padding)을 사용한다. 패딩이란 입력 데이터의 모서리 주변 부분을 특정 데이터로 채우는 것을 의미하는데, 주로 0의 값을 할당하는 방법(zero-padding)을 사용한다. 스트라이드(stride)는 입력 데이터에서 필터의 윈도우가 얼마의 간격으로 이동할 것인가 결정하는 하이퍼 파라미터를 의미한다. CNN의 일반적인 알고리즘에서 2차원 데이터의 공간 축소를 위해 풀링(pooling)이 사용되는데 풀링은 해당 필터의 픽셀 데이터 중 최대값이나 평균값으로 이미지 정보를 재구성 한다. 이미지 분석에서는 일반적으로 평균값 풀링(average pooling)보다는 최대값 풀링(max pooling)이 사용된다. 최종적으로 여러 단계의 합성곱 계층 연산 및 풀링을 통해 완전 연결 계층 연산으로 분류를 수행한다.

2.2 하이웨이 네트워크

하이웨이 네트워크는 (1)의 식과 같이 LSTM(Long Short-Term Memory)의 게이팅 유닛(gating unit)을 활용하는 방식으로 입력 데이터의 값과 모델링의 학습 결과 값을 연결하는 방식이다(Srivastava et al., 2015). 하이웨이 네트워크 레이어(highway network layer)에서는 입력 데이터의 신호를 그대로 분류에 활용할 것인지, 분류 모델링의 네트워크를 통해 학습하여 변환하여 최종 결과 값에 반영할 것인지 결정한다. 이와 같은 절차를 통해 딥러닝 네트워크에서 히든 레이어(hidden layer)가 많을수록 입력 데이터가 희미해지는 그래디언트 소

실(gradient vanishing) 문제를 해결한다.

$$y = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_T) \quad (1)$$

(1)의 식에서 x는 입력 데이터, y는 출력 데이터, H는 아핀 변환(affine transform) 비선형 활성화 함수, T는 비선형 변환 게이트(transform gate), C는 비선형 통과 게이트(carry gate)를 의미한다. $C = T - 1$에 해당하며 T 게이트가 1일 경우 x값이 분류 모델의 네트워크를 통해 변환하고, 0일 경우 그대로 모델 네트워크를 거치지 않고 통과하여 y값으로 산출된다.

2.3 악성 댓글 분류

악성 댓글 분류에 관한 연구는 일반적으로 온라인 뉴스기사, 블로그, SNS 등의 댓글들을 대상으로 했다. 과거 단어 사전을 기반으로 하는 연구에서 최근에는 머신 러닝 기술을 활용하여 문장의 맥락까지 고려할 수 있는 방법으로 발전했다. 국내 연구의 경우 배민영 등(2009)은 댓글의 악성 여부를 분류하기 위해 1,300개의 한국어 댓글을 n-gram으로 나누고 토픽 시그니처(topic signiture) 기술을 활용하여 특성을 추출하는 방식의 분류 모델을 연구했다. 특성 추출 후 SVM(Support Vector Machine)으로 악성 댓글을 분류했을 때 f1 점수 기준 약 88%로 기존 연구에 비교적 우수한 성능을 보였다고 한다. 김현정 등(2011)은 FFP(Feature Frequency Profile) 특성 추출 기법을 기반으로 1,700개 한국어 댓글 데이터에 대해서 SVM 및 Random Forest 분류를 통해 모델의 성능을 개선했다. 결과적으로 FFP기반 SVM 분류 모델링으로 약 79%의 성능이 나타났다고 한다. 홍진주 등(2016)은 악성 댓글의 특성상 단어들의 변형이 많다는 점을 고려하여 600개에 해당하는 댓글의 학습에 감성 분석 기반 단어 사전 및 SVM 분류 모델링을 적용한 결과로 분류 모델의 성능을 개선했다. 결과적으로 f1 점수 기준 87%를 달성했다. Moon et al(2020)은 10,000개에 해당하는 악성 댓글을 차별과 혐오에 해당하는 두 가지 분류 기준으로 세 가지 클래스를 분류하여 데이터 셋을 구축했다. 이를 딥러닝 기술을 활용하여 단어 기반이 아니라 문장의 맥락을 고려하여 분석하고자 했다. 분석한 결과 KoBERT(Korean Bidirectional Encoder Representations from Transformer)를 적용한 분류 모델 성능이 차별 기준 데이터에서 세 가지 클래스에 대해 f1 점수 기준 63.3%로 가장 우수하게 나타났으며 혐오 기준에서는 제대로 분류가 되지 않음을 발견했다. 성지석과 임희석(2020)은 17,000개의 영어 댓글 데이터에 대해 BERT, GCN(Graph Convolution Network), GAT(Graph Attention Network)와 같은 그래프 구조를 이용한 악성 댓글 분류 시스템을 설계했다. 결과적으로 BERT의 분석 결과에 비해 분류 모델링 시간을 대폭 축소했다고 하며 GCN 모델링을 적용했을 때 f1 점수 기준 49%를 달성했다.

국외 연구의 경우 Li(2018)는 악성 댓글 분류 문제를 해결하기 위해 word2vec의 skip-gram 임베딩과 BiLSTM(Bidirectional Long Short-Term Memory)을 활용했다. 또한 악성 댓글의 세부 항목을 분류하여 각각에 대한 모델링을

구축했으며 전반적으로 f1 점수 기준 70~80%가 측정되지만 매우 공격적인 표현(Severe Toxic) 항목과 공격적인 표현(Toxic) 항목을 봤을 때 Moon et al(2020)과 같이 공격성의 정도를 구분하여 학습하지는 못하는 것으로 나타났다. Srivastava et al(2018)은 일반적인 웹사이트에 환경에 적용하기 위해서 경량화된 딥러닝 기반 캡슐 네트워크(capsule network) 모델링 기술을 제안했다. 제안된 캡슐 네트워크 악성 댓글 분류 모델링의 경우 f1 점수 기준 성능이 63%로 나타났다. Georgakopoulos(2018)은 CNN 기법을 적용하여 악성 댓글 분류 성능을 f1 점수 기준 91.7%로 상당한 수준으로 성능을 개선했다. Carta et al(2019)는 다중 라벨(label)에 대해 제안된 지도 학습 기반 워드 임베딩 기법과 다양한 워드 임베딩(word embedding) 기술을 적용하여 6개의 악성 댓글의 세부 항목 클래스에 대해 분류 했다. 결과적으로 제안된 워드 임베딩 기법이 AUC(Area under the Curve) 기준 85~88%에 해당하는 결과를 달성했다.

국외 연구의 경우 딥러닝 모델링 기술을 기반으로 문장의 맥락을 고려하여 악성 댓글을 분류하거나 악성 댓글의 세부 항목까지 분류하는 연구들이 최근 등장하고 있다. 하지만 한국어 대상 악성 댓글 분류 연구들은 사전 기반 분석으로 문장의 맥락을 고려하지 못하고 차별 및 혐오 표현에 대한 세부 항목을 포함하지 않으며 신조어를 처리하기 어렵다는 문제점이 있다(배민영 등, 2009; 김현정 등, 2011; 홍진주 등, 2016). 최근에는 한국어 대상으로도 국외 연구 동향과 같이 딥러닝 기술을 활용하여 문장의 맥락을 고려하는 악성 댓글 분류 모델링을 연구하거나 세부 항목을 고려하는 연구들을 수행하고 있긴 하지만 모델의 성능적인 부분에서 알고리즘의 고도화가 필요한 상황이다(Moon et al., 2020).

Ⅲ. 연구 방법론

3.1 데이터

본 연구는 Moon et al(2020)에서 어노테이션(annotation)한 <표 1>과 같은 한국어 온라인 뉴스 댓글 데이터 셋 샘플을 활용한다. 이 데이터 셋 샘플은 2020 인공지능 온라인 경진대회의 18번 과제인 '인터넷 악성 댓글 필터링을 위한 분류 모델 개발'에서 제공되었다(AI Challenge, 2020). 데이터 샘플은 총 8,878개이고 학습(training) 데이터, 검증(validation) 데이터, 테스트(test) 데이터로 구분하여 각각 7,867, 500, 511개로 할당한 것이다. 악성 댓글 데이터의 라벨링(labelling)은 두 가지 분류 기준으로 이루어지는데, 차별(Bias)과 혐오(Hate)가 이에 해당한다. 차별과 혐오는 각각 세 가지 클래스로 분류하며 차별은 다시 성적인 차별(Gender), 기타 차별(Others), 차별 없음(None)으로, 혐오는 댓글의 공격성에 따라서 혐오(Hate), 공격성(Offensive), 혐오 없음(None)으로 나누어진다. 차별적 기준의 Gender 클래스는 성적인 차별이 포함되는 차별적 표현, Others는 성별 외 인종, 출신 지역, 피부색, 종교, 장애, 직업 등에 대한 편견이 포함되는 차별적 표현, None은 편견이 존재하지 않는 댓글을 의미한다. 혐오적 기준의 Hate 클래스는 대상을 심하게 비난하거나 인신공격을 하여 정신적인 고통을 유발할 수 있는

혐오적 표현, Offensive는 Hate의 혐오적 표현 수준에는 미치지 않지만 공격적이고 무례한 표현, None은 혐오적인 표현이 존재하지 않는 댓글을 의미한다. 데이터의 형식은 인터넷 뉴스 기사의 제목과 댓글, 그리고 차별과 혐오 클래스 분류들로 구성되어 있다. 모델링 과정에서 뉴스 기사 제목이 성능을 떨어트리는 것으로 확인되어 본 분석에는 제거했다. 학습 및 검증 데이터의 각 클래스 분포는 <표 2>와 같다. 전체적으로 None 클래스의 수가 가장 많으며 차별 기준에서는 클래스 분포 편차가 심한 편이다. 통합 클래스의 경우 Gender None과 Others None의 경우 클래스 데이터의 불균형이 있기 때문에 샘플링 기법이나 클래스 가중치 조정과 같은 해결 방안이 필요하다. 테스트 데이터의 라벨은 2020 온라인 인공지능 경진대회의 규칙상 조작을 피하기 위해 공개되지 않는다.

3.2 모델링

본 연구의 악성 댓글 분류 모델은 <그림 2>, <그림 3>과 같이 성능 개선을 위해 CNN, 하이

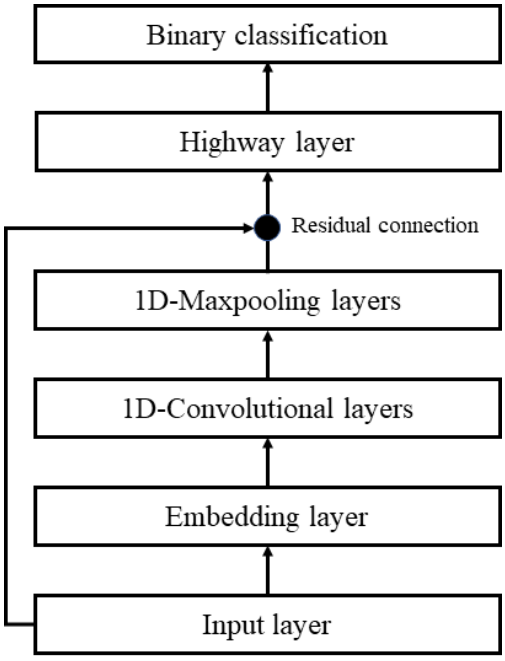
<표 1> 악성 댓글 데이터 예시

Comment	Bias	Hate
1,2화 어설펐는데 3,4화 지나서부터는 갈수록 너무 재밌던데	None	None
12월이나 1월이더 결혼성수기지 5월은 여름이라 비수기에속한다	None	None
15년생인 울아들은 유모차 안타는데	None	None
2년 사귀었으면 ... 서로 온몸의 점 위치까지 다알 듯 ㅋㅋㅋ	Others	Offensive
40살에 폐경에 다리까지 노답이라이 ㅋㅋㅋ	Gender	Hate
니같은 개돼지가있어서 개한민국이 엉망진창인거야. 주제를알아야지. 일베충만큼 개돼지인문베충나라문화방송말아먹는쓰레기들	Others	Hate

<표 2> 데이터 클래스 분포

분류 기준	클래스	수	통합 클래스	수
차별	Gender	1299	Gender Hate	833
	Others	1578	Gender None	83
	None	5490	Gender Offensive	383
혐오	Hate	2033	None hate	557
	Offensive	2688	None None	3422
	None	3646	None Offensive	1511
			Others Hate	643
			Others None	141
			Others Offensive	794

웨이 네트워크, 사전 외 처리 기술에 해당하는 OOV(Out of Vocabulary) 사전학습 임베딩을 활용했다. 본 연구에서 활용한 CNN 기법은 텍스트 분류 연구에서 우수한 성능을 나타내는 Kim-CNN 기법을 활용했다. Kim-CNN은 한 문장에 포함되는 단어들을 벡터로 임베딩해서 여러 너비(width)값의 다중 필터(multiple filter)를 가지는 합성곱 계층을 구성한다(Kim, 2014). 이를 최대오버타임(max-over-time) 풀링 후 완전 연결 계층을 통해 분류 값을 산출한다. 다중 필터의 너비 값을 다양하게 지 정하여 분류 값을 산출할 때 고려하는 단어의 수를 변형할 수 있다.

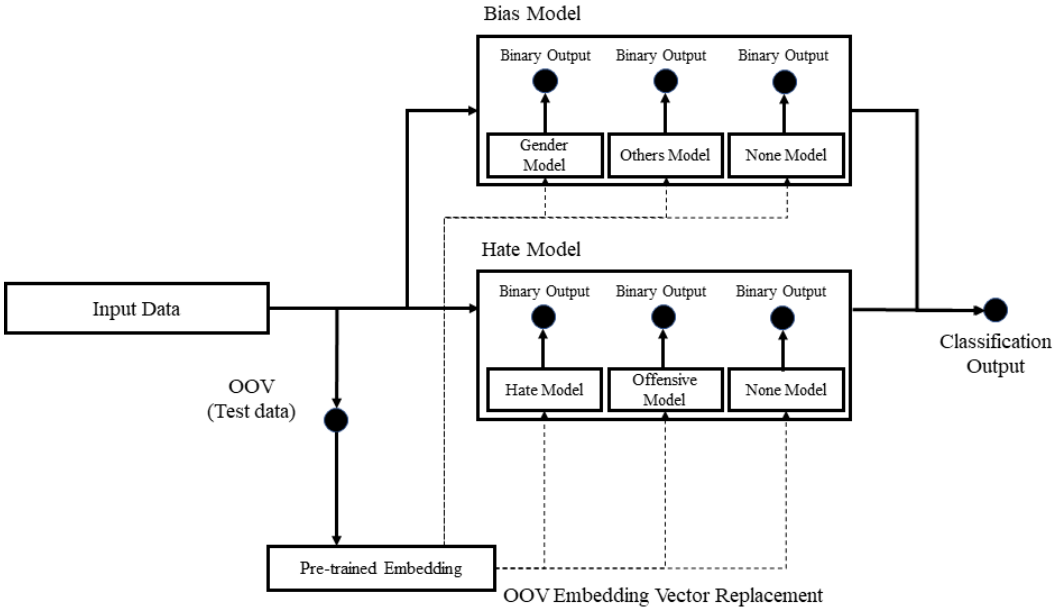


<그림 2> 이진 분류 모델링 설계

본 연구에서는 <그림 2>처럼 입력 레이어(input layer)에서 텍스트 데이터를 수치화하여 모델링에 적용하기 위해 단어들을 임의의 숫자로 변환하여 임베딩 레이어(embedding layer)를 구축했다. 이 임베딩 레이어를 Kim-CNN 기법으로 여러 너비 값을 가지는 필터로 이루어진 1차원 컨볼루셔널 레이어(1D-convolutinal layers)로 분할하고, 차원 축소를 위해 이를 1차원 최대 풀링 레이어(1D-maxpooling layers)로 구성했다. 이 때 각 컨볼루셔널 레이어와 풀링 레이어는 각기 다른 필터 너비 값을 가지는 총 5개의 레이어를 합친 것이다. 각 레이어의 필터 너비는 1에서 5까지의 값을 부여했으며 필터 너비에 따라 분류 값을 산출할 때 고려하는 단어 수를 할당하여 학습한다.

그 다음 CNN 모델링 학습 값과 입력 레이어와의 잔차 연결(residual connection)을 통해 하이웨이 네트워크 레이어를 구성했다. 본 연구에 활용된 하이웨이 네트워크 기법은 CNN 모델링의 1차원 최대 풀링 레이어의 학습 결과 값과 입력 레이어의 입력 값을 기반으로 최종적인 이진 분류 값 산출에 가중치를 부여하는 방식이다. 하이웨이 네트워크 레이어는 입력 레이어의 신호를 그대로 반영할지, 아니면 모델링의 네트워크를 통해 변환할지 결정하는 역할을 수행한다. 본 연구에서는 CNN 모델링에 하이웨이 네트워크를 접목했을 때 분류 모델의 성능이 개선되었다.

OOV에 해당하는 테스트 데이터에만 있는 유니크(unique) 단어들은 벡터 가중치의 값이 학습 되지 않기 때문에 테스트 데이터 예측 시 성능 저하가 나타날 수 있다. 따라서 <그림 3>과 같이 OOV 처리를 위해 입력 데이터의 말뭉치(corpus)에 속하는 단어들을 Word2Vec 기법을 통해 사전학습(pre-training)하여 OOV와 벡



<그림 3> 분류 모델링 설계

터 공간상의 유사한 단어들을 추출한다. 본 연구에서는 추출한 유사 단어들의 모델 학습 과정에서 생성된 벡터 가중치 값의 평균을 테스트 데이터의 예측에 활용했다. Word2Vec을 통한 단어 임베딩(word embedding)은 유사 단어들을 추출하는 것에 활용하고 실제 학습 모델에 들어가는 단어 임베딩은 형태소 단위의 토큰(token)에 임의의 숫자를 부여하는 방식을 활용했다. 본 연구에 활용된 OOV 처리 방식은 분류 모델의 성능을 개선했다.

여기서 Word2Vec 기법이란 여러 문장 내의 중심 단어들을 기준으로 그 주위에 어떤 단어들이 출현하는 지 학습을 통해 예측하는 것으로 단어들을 벡터 공간에 배치하여 단어들 간의 상호 유사도를 추출할 수 있는 방식이다(Mikolov et al., 2013). 일반적으로 Word2Vec 기법은 사전학습된 외부 데이터를 활용하여 미세 조정(fine tuning) 후 임베딩하는 방식으로 OOV 등의 문제를 처리하지만 2020 온라인 인공지능 경진대회의 규칙상 외부 데이터를 활용할 수 없었다. 따라서 이를 보완하기 위해 말뭉치의 모든 단어들을 Word2Vec으로 학습하여 임베딩으로 활용한 것이다.

본 연구의 연구 모형에서 목표는 단순히 악성 댓글에 대한 이진 분류 모델링을 하는 것이 아니라 악성 댓글의 세부 항목까지 분석하기 위해 다중 레이블(multi-label) 분류 모델링을 구현하는 것이다. 이에 차별 및 혐오 기준에 해당하는 각각 3개의 클래스로 시그모이드 이진 분류(sigmoid binary classification)를 하는 방식으로 구축한 후, 데이터 불균형을 해결하기 위해 각각의 모델에 대해 클래스 가중치(class weight)를 부여하고, 세밀하게 파라미터(parameter)를 조정했다. 각각의 차별 및 혐오

기준에 따라 클래스에 대해 3개씩 소프트맥스(softmax)로 분류하는 방식이 아니라 시그모이드 이진 분류를 하여 통합하는 방식을 채택한 이유는 다중 레이블 분류 모델에서 레이블이 배타적이지 않은 경우 이진 분류 방식이 효과적이라는 선행 연구의 방법론을 따른 것이다(Redmon et al., 2018). 이러한 방법론은 이미지 분석 분야에서 활용되고 있으며 본 연구의 경우도 각 클래스에 대한 통합 결과를 산출 시 차별 및 혐오 기준의 클래스가 서로 배타적이지 않은 경우기 때문에 모델의 성능이 개선된 것으로 판단된다.

또한 일반적으로 데이터의 클래스 불균형의 문제는 오버 샘플링(over sampling)이나 네거티브 샘플링(negative sampling) 기법을 활용하지만 이러한 기법들이 입력 데이터의 수가 적은 본 연구의 모델링에서는 성능을 저하시켰음을 확인했다. 이는 데이터의 수가 적어지기 때문에 네거티브 샘플링의 경우 부적합하고 오버 샘플링의 경우 단순히 적은 수의 클래스에 대한 보충 시 과적합 문제가 일어날 수 있기 때문이라고 판단된다. 따라서 각각의 이진 분류 모델에 대해 파라미터에 해당하는 클래스 가중치를 데이터 불균형에 따라 세부 조정한 결과 분류 모델의 편향성을 보완하여 성능을 개선했다. 그 외 세부 파라미터의 경우 컨볼루셔널 및 풀링 레이어의 수, 신경망의 뉴런의 개수, 드롭아웃(dropout) 등의 수치를 모델 성능에 따라서 최적화 했다. 최종적으로 산출된 각각의 이진 분류 모델링의 검증 및 테스트 평가의 결과 값을 하나로 통합하여 가중 평균 f1 점수로 나타냈다.

Ⅳ. 연구 결과

연구 결과 모델의 성능이 <표 3>과 같이 가중 평균 f1 점수 기준 테스트 데이터에서 차별 및 혐오 기준 클래스 총합 67.49%로 나타났다. f1 점수란 (2)의 식과 같이 연산되며 정밀도(precision)와 재현율(recall)의 조화평균 값으로 나타난다. f1 점수가 조화평균 값이기 때문에 정밀도와 재현율 중 더 적은 값에 큰 가중치가 부여된다. 클래스가 불균형한 상태일 때 모델의 성능을 정확하게 평가할 수 있으며 하나의 백분율 수치로 나타난다. 정밀도는 (3)의 식처럼 TP(True Positive)값을 TP와 FP(False Positive)를 더한 값으로 나눈 값이며 예측 값 중에서 얼마나 정확하게 분류했는가에 대한 수치를 의미한다. 재현율은 TP값을 TP와 FN(False Negative)을 더한 값으로 나눈 값이며 실제 값 중에서 얼마나 정확하게 분류했는지에 대한 수치를 의미한다. 본 연구의 모델 평가는 테스트 데이터에 대해 차별 기준 모델과 혐오 기준 모델에서 산출된 가중 평균 f1 점수 값의 평균으로 산정했다. 가중 평균 f1 점수 값은 각 클래스에 해당하는 샘플 수를 총 샘플 수로 나누어 가중치로 환산하여 각 클래스에 대한 f1 점수를 곱하여 합산하는 방식이다. 최종 통합 f1 점수는 차별 기준 클래스에 대한 가중 평균 f1 점수와 혐오 기준 클래스에 대한 가중 평균 f1 점수를 합하여 2로 나눈 값이다.

$$f1\text{점수} = \frac{2 \times \text{정밀도} \times \text{재현율}}{\text{정밀도} + \text{재현율}} \quad (2)$$

$$\text{정밀도} = \frac{TP}{TP + FP} \quad (3)$$

$$재현율 = \frac{TP}{TP + FN} \quad (4)$$

$$가중평균\ f1\ 점수 = \frac{\sum_{i}^{n} 클래스 i\ 샘플수 \times 클래스 i\ f1 점수}{전체\ 샘플수} \quad (5)$$

모델 평가 결과 검증 데이터에서 가중 평균 값이 69.8%으로 나온 모델이 테스트 데이터에서도 f1 점수 값이 67.49%로 안정적인 성능을 나타내는 것을 확인했다. 본 연구의 결과는 같은 어노테이션 기준을 가지는 한국어 텍스트 데이터를 기반한 Moon et al(2020)의 모델 테스트 결과와 비교했을 때 크게 상회하는 수준이다. 해당 연구는 본 연구 문제의 기반이 되는 2020 온라인 인공지능 경진대회의 18번 과제에 관련된 것으로 모델링에 사용된 약 10,000개의 데이터는 본 연구의 데이터와 상당 부분 같다고 볼 수 있다. Moon et al(2020)에서는 해당 데이터 셋을 분석했을 때 차별 기준 클래스의 경우 가중 평균 f1 점수가 63.3%, 혐오 기준 클래스의 경우 제대로 분류되지 않았다고 한다. 본 연구의 모델에서는 검증 데이터에서 차별 기준 클래스의 f1 점수가 77.34%, 혐오 기준 클래스에서는 62.25%로 보다 높은 성능이 나타났다. 혐오 기준 클래스의 경우 클래스 분류 기준이 혐오 정도의 차이인 이유로 제대로 개념적 분류가 되지 않아 모델 성능을 개선하는 데 어려움이 있었다. 본 연구는 2020 인공지능 경진대회의 규칙을 기반으로 했기 때문에 테스트 데이터에서 차별 및 혐오 기준 클래스의 평가 결과를 따로 확인할 수 없었다(AI Challenge, 2020). 이는 경진대회의 절차상으로 테스트 데이터의 라벨이 없는 상태에서 학습 및 검증 데이터를 기반으로 한 모델로 해당 시스템에서 자체적으로 평가를 하기 때문이다. 하지만 본 모델의 검증 및 테스트 결과를 봤을 때 기존 한국어 대상 선행연구와 비교적 충분히 우수한 성능을 나타냄을 알 수 있었다.

<표 3> 학습 모델의 가중 평균 f1 점수

클래스	가중 평균 f1 점수
Bias (Validation)	0.7734
Hate (Validation)	0.6225
Bias + Hate (Validation)	0.6980
Bias + Hate (Test)	0.6749

국외 영어 대상 논문의 경우 일반적으로 Kaggle Competition의 'Toxic Comment Classification Challenge'의 데이터 셋을 기반으로 분류 모델링을 연구했으며 'toxic', 'severe toxic', 'obscene', 'threat', 'insult', 'identity hate'에 해당하는 세부 항목을 가지고 있다(Kaggle, 2018). Li(2018)는 본 연구와 같이 악성 댓글의 세부 항목을 분류하여 70~80%의 f1 점수를 항목 별로 달성했으나 공격적인 표현의 부분에서는 제대로 결과 값이 나오지 않았다. Srivastava et al(2018)의 논문에서는 악성 댓글에 대한 이진 분류 모델링 시 f1 점수가 약 63%로 나타났다. Carta et al(2019)는 제안된 워드 임베딩을 활용했을 때 기본 분류 모델링 분석으로 AUC기준 85~88%의 성능을 보였으나 f1 점수를 측정하지 않았기 때문에 편향된 클래스에 대해서 어떻게 분류했는지 나타나지 않았다. Georgakopoulos(2018)에서는 악성 댓글 이진 분류 CNN 모델링의 결과 값이 f1 점수 기준

91.7%로 관련 연구 중 가장 우수한 수준으로 나타났다.

선행 연구들의 한국어 대상 딥러닝 기반 악성 댓글 분류 모델의 성능 결과를 살펴봤을 때, 악성 댓글의 세부 항목을 고려한 연구 중에서 본 연구의 모델 성능이 비교적 우수하다는 점을 알 수 있었다. 특히 국외 연구에서도 제대로 분류 모델 성능 결과를 달성하지 못했던 공격적인 표현 부분에서의 f1 점수 결과를 약 62%까지 달성했다는 점, 한국어 대상 딥러닝 기반 악성 댓글 분류 연구 중에서의 성능이 크게 우수한 수준이라는 것을 확인했다. 실제로 <표 4>와 같이 본 연구의 모델링을 기반으로 한 예측 사례를 살펴봤을 때, 악성 댓글에 대해 준수하게 분류하는 모습을 보였다.

기존 다양한 텍스트 분류 및 시퀀스 분석 분야에서 BERT(Bidirectional Encoder Representations from Transformers)가 가장 우수한 모델 성능을 보이지만, 악성 댓글 분류 분야의 연구에서는 아직까지 고도화된 CNN 모델링을 활용한 연구들의 모델 성능이 우수한 것으로 나타났다(Georgakopoulos, 2018). 이는 온라인상의 다양한 댓글들이 대체적으로 텍스트의 길이가 짧고 변칙적이라는 특성을 가지고 있기 때문이라고 판단된다. 또한 BERT 모델의 특성상 수많은 데이터를 활용하고 실무적으로 적용 시 모델의 분류에 걸리는 시간이 오래 걸릴 수 있기 때문에 경량화에 대한 문제점이 있다. 본 연구에서도 이와 같은 요소와 함께 2020 인공지능 온라인 경진대회의 규칙으로 BERT의 사전학습 데이터를 활용할 수 없는 상황이라 BERT 기법보다는 CNN 기반 모델링이 적합했다.

<표 4> 악성 댓글 분류 예측 예시

Comment	Bias	Hate
둘다 넘 좋다~행복하세요	None	None
근데 만원이하는 현금결제만 하라고 써놓은집 우리나라에 엄청 많은데	None	None
누군데 얘네?	None	Offensive
페미들 ㄹㅇ 토나온다	Gender	Hate
그래 자랑이다. 개독이 얼마나 무서운데 얼마나 갈지 함 두고보자.	Others	Offensive
아 시발 더러워 ㅡㅡㅋ 제발 늙은것들은 젊은 사람 탐하지마라 남녀 모두 ㅡㅡ 진짜 더러워	Others	Hate

Ⅴ. 결 론

본 논문은 하이웨이 네트워크 기반 CNN, OOV 사전학습 임베딩 방식으로 모델의 성능을 개선하여 기존 연구에 비해 비교적 높은 성능을 나타내는 모델링 방법론을 연구했다(Moon et al., 2020). 모델 테스트 평가 결과 가중 평균 f1 점수 기준 67.49%로 같은 분류 기준을 가지는 Moon et al(2020)의 모델보다 크게 상회하는 수준으로 나타났다. 본 논문의 시사점은 다음과 같다.

본 논문은 한글을 대상으로 악성 댓글을 효과적으로 분류할 수 있는 방법론을 제시한다. 기존의 한글 대상 악성 댓글 분류 연구는 대부분 단어 기반이고, 데이터가 적으며, 문장의 맥락을 고려하지 못한다는 문제점이 있다(배민영 등, 2009; 김현정 등, 2011; 홍진주 등, 2016). 본 연구에 활용된 데이터 셋의 기반이 되는 Moon et al. 2020)의 경우 비슷한 맥락으로 한국어 대상 딥러닝 기반 악성 댓글 분류 모델링

연구지만 모델 성능의 고도화가 필요한 상황이다. 따라서 본 연구는 하이웨이 네트워크 기반 CNN 기법을 활용하여 모델의 성능을 개선한 악성 댓글 분류 모델을 개발했다. 본 연구의 모델은 2020 인공지능 경진대회에서 활용한 것으로 대회 규칙상 외부 데이터를 활용할 수 없었지만 기존 연구 대비 비교적 우수한 성능이 나타났으며 악성 댓글 분류 과제에서 1등을 달성했다(AI Challenge, 2020). 외부 데이터를 활용할 경우 모델의 성능을 개선할 수 있을 것으로 기대된다.

본 연구에서는 OOV에 대해 사전학습 임베딩을 하는 방식으로 기존 OOV 처리 문제를 해결했으며, 악성 댓글의 세부적 유형을 효과적으로 분류할 수 있는 모델을 개발했다. 본 연구에서 활용된 OOV 처리는 실무적으로 활용할 때 Word2Vec이나 FastText와 같은 사전학습 임베딩을 미세 조정을 통해 활용한다면 분류 모델의 성능을 개선할 수 있을 것으로 기대한다. 이때 미세 조정에 활용되는 텍스트 데이터는 라벨이 없는 댓글 데이터도 활용할 수 있기 때문에 수월하게 데이터베이스를 구축할 수 있다. 이에 따라서 신조어나 특이 단어에 대해서도 모델에서 학습된 가중치를 활용하여 대처할 수 있다.

이와 같은 악성 댓글 분류 방법론을 통해 실무적으로는 인터넷 뉴스, SNS, 온라인 동영상 플랫폼 등의 악성 댓글 통제뿐만 아니라 공공서비스에 적용하여 민원 및 소통창구 등의 텍스트에서도 차별 및 혐오 표현을 탐지하여 필터링(filtering)할 수 있을 것으로 기대된다. 또한 최근 1인 미디어 및 라이브 커머스(live commerce) 시장이 급부상 하고 있는데, 실시간 채팅에 대한 차별 및 혐오 표현 문제도 본 연구의 방법론을 적용한다면 효과적으로 통제할 수 있을 것으로 예상된다. 결론적으로 인터넷 공간에서 차별 및 혐오적 발언으로 타인에게 고통을 주는 악성 댓글을 효과적으로 통제함으로써 건전한 인터넷 문화에 기여할 수 있을 것이다.

본 논문의 한계점은 다음과 같다. 8,878개의 샘플 데이터를 활용했기 때문에 클래스 당 데이터의 개수가 부족했다. 향후 연구 방향으로 더 큰 데이터베이스(database)를 구축하여 보다 우수한 성능을 가지는 모델을 개발할 수 있을 것으로 기대된다. 본 연구는 Moon et al(2020)의 어노테이션 기준을 가지고 연구를 수행했지만 차별 및 혐오 기준의 클래스 분류가 모호하거나 더 세부적인 클래스 분류가 필요할 수 있다. 이는 데이터베이스 구축 시 차별 기준의 항목을 보다 세분화하고 혐오 기준의 경우 혐오의 정도를 점수를 매기는 방식으로 개선할 수 있다. 또한, 본 연구는 외부 데이터를 활용하지 않았기 때문에 Word2Vec, GloVe, FastText 등의 사전학습 임베딩이나, BERT 기법을 활용할 경우 성능 개선의 여지가 존재한다.

참고문헌

경찰청, “전체 사이버범죄 발생 및 검거 현황,” 2020, https://www.police.go.kr/www/open/publice/publice0204.jsp

과학기술정보통신부, “2020 AI challenge,” 2020, http://www.aichallenge.or.kr/main/main.do

과학기술정보통신부, “2019 인터넷이용실태조

사 결과 발표," 2020.

김현귀, "인터넷 실명제의 도입과 헌법재판소 결정", 헌법판례연구, 제14권, 2013, pp: 157-192.

김현정, 윤영미, 이병문, "향상된 FFP(Feature Frequency Profile)를 활용한 악성 댓글의 판별시스템," 한국정보기술학회논문지, 제9권, 1호, 2011, pp: 207-216.

배민영, 은지현, 장두성, 차정원, "지지 벡터 기계와 토픽 시그너처를 이용한 댓글 분류 시스템: 언어에 독립적인 댓글 분류 시스템," 한국 HCI 학회 학술대회, 2009, pp: 263-266.

성지석, 임희석, "그래프 구조를 이용한 악성 댓글 분류 시스템 설계 및 구현," 한국융합학회논문지, 제11권, 6호, 2020, pp. 23-28.

양낙영, 김성근, 강주영, "텍스트 마이닝 방법론과 메신저 UI를 활용한 융합연구 촉진을 위한 연구자 및 연구 분야 추천 시스템의 제안," 정보시스템연구, 제27권, 4호, 2018, pp. 71-96.

정건용, 윤승식, 강주영, "재정정보 활용을 위한 텍스트 마이닝 기반 회계용어 형태소 분석기 구축. 정보시스템연구," 제28권, 4호, 2019, pp. 155-174.

홍진주, 김세한, 박제원, 최재현, "감성분석과 SVM을 이용한 인터넷 악성 댓글 탐지 기법," 한국정보통신학회논문지, 제20권, 2호, 2016, pp: 260-267.

(주)한국리서치, "[기획] 악성 댓글, 이대로 괜찮은가", https://hrcopinion.co.kr/archives/14589, 2020.

Carta, S., Corriga, A., Mulas, R., Recupero, D. R., and Saia, R.. "A Supervised Multi-class Multi-label Word Embeddings Approach for Toxic Comment Classification," *Paper presented at the KDIR*, 2019.

Georgakopoulos, S. V., Sotiris K. T., Aristidis, G. V., and Vassilis, P. P., "Convolutional Neural Networks for Toxic Comment Classification," *Paper presented at the Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, 2018.

Kaggle Competition, Toxic Comment Classification Challenge, 2017, Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

Kim, Y., "Convolutional Neural Networks for Sentence Classification," arXiv preprint arXiv:1408.5882, 2014.

Park, K., and Ha, S., "Customer Service Evaluation based on Online Text Analytics: Sentiment Analysis and Structural Topic Modeling," *Korea Association Information Systems*(정보시스템연구), Vol. 26, No. 4, 2017, pp. 327-353.

LeCun, Y., Bernhard, E. B., John, S. D., Donnie H., Richard, E. H., Wayne, E. H., and Lawrence, D. J., "Handwritten Digit Recognition with a Back-Propagation Network," *Paper presented at the*

Advances in neural information processing systems, 1990.

Li, S., "Application of Recurrent Neural Networks in Toxic Comment Classification," UCLA Master's Thesis, 2018.

Mikolov, T., Kai, C., Greg, C., and Jeffrey D., "Efficient Estimation of Word Representations in Vector Space", arXiv preprint arXiv:1301.3781, 2013.

Moon, J., Cho, I., and Lee, J., "Beep! Korean Corpus of Online News Comments for Toxic Speech Detection," arXiv preprint arXiv:2005.12503, 2020.

Reuters Institute, "Digital News Report 2020," 2020.

Srivastava, R. K., Klaus G., and Jürgen S., "Highway Networks", arXiv preprint arXiv:1505.00387, 2015.

Srivastava, S., Prerna K., and Vartika T., "Identifying Aggression and Toxicity in Comments Using Capsule Network," *Paper presented at the Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 2018.

이 현 상 (Lee, Hyun-Sang)

경북대학교에서 경영학 학사와 석사 학위를 취득했으며 현재 경북대학교 경영학부 박사과정에 재학 중이다. 주요 관심 분야는 머신러닝 및 딥러닝 기반 예측 모델, 시계열 분석, 그리고 텍스트 분류 및 마이닝 등이다.

이 희 준 (Lee, Hee-Jun)

2019년 계명대학교에서 경영정보학 전공과 비즈니스데이터분석 부전공으로 학사학위를 취득하고 계명대학교 경영정보학과에서 석사과정에 재학 중이다. 연구분야는 딥러닝, 머신러닝, 추천시스템이다.

오 세 환 (Oh, Se-Hwan)

현재 경북대학교 경영학부에서 조교수로 재직 중이다. 서울대학교 경제학부(학사)를 졸업했으며 카네기멜론대에서 e-비즈니스 석사, 서울대학교에서 경영학(경영정보) 박사 학위를 받았다. International Journal of Mobile Communications, Internet Research, Journal of Electronic Commerce Research 등에 논문을 게재했으며 주요 연구 관심분야는 공유경제, 전자상거래, 온라인 구전 등이다.

<Abstract>

A Study on the Toxic Comments Classification Using CNN Modeling with Highway Network and OOV Process

Lee, Hyun-Sang · Lee, Hee-Jun · Oh, Se-Hwan

Purpose

Recently, various issues related to toxic comments on web portal sites and SNS are becoming a major social problem. Toxic comments can threaten Internet users in the type of defamation, personal attacks, and invasion of privacy. Over past few years, academia and industry have been conducting research in various ways to solve this problem. The purpose of this study is to develop the deep learning modeling for toxic comments classification.

Design/methodology/approach

This study analyzed 7,878 internet news comments through CNN classification modeling based on Highway Network and OOV process.

Findings

The bias and hate expressions of toxic comments were classified into three classes, and achieved 67.49% of the weighted f1 score. In terms of weighted f1 score performance level, this was superior to approximate 50~60% of the previous studies.

Keyword: deep learning, Highway Network, CNN, OOV, toxic comments

* 이 논문은 2020년 8월 31일 접수, 2020년 9월 6일 1차 심사, 2020년 9월 11일 게재 확정되었습니다.