JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 12: 1880~1889, Dec. 2022

어텐션임베딩과 다채널 CNN 기반 반시민성 검출 알고리즘

박윤정[1] · 이세영[2] · 금희조[3*]

Detection of Incivility based on Attention-embedding and multi-channel CNN

Youn-Jung Park[1] · Se-Young Lee[2] · Hee-Jo Keum[3*]

[1]Senior Researcher, Global Convergence Contents Research Center, Sungkyunkwan University, Seoul, 03063 Korea
[2]Associate Professor, Department of media communication, Sungkyunkwan University, Seoul, 03063 Korea
[3*]Professor, Department of media communication, Sungkyunkwan University, Seoul, 03063 Korea

요 약

온라인 포털 플랫폼은 뉴스 기사와 온라인 댓글을 제공하고 있으나, 온라인 댓글의 익명성은 반시민적 표현을 증가시켜 사회적 문제점으로 간주되고 있다. 댓글의 반시민성 검출 연구가 많이 이루어진 국외와 달리, 국내에서는 비시민성을 세분화한 한국어 데이터셋이 구현되지 않아 심도있는 연구가 이루어지지 못하였다. 본 연구에서는 댓글의 반시민성에 대한 라벨링을 총 13가지 항목으로 시행하였으며 반시민적 표현으로 요약하였다. 또한 어텐션 알고리즘을 이중으로 적용하여 임베딩 벡터를 추출하였고 이후 2-d CNN으로 반시민성 항목을 분류하였다. 그 결과, 제안한 알고리즘이 무례한 호칭 및 공격적 어조 등의 반시민성 검출에 유용하다는 것을 보여주었다. 본 연구는 민주적 담론을 저해하는 반시민적 댓글들을 탐지함으로써 건전한 온라인 댓글 문화 형성에 기여할 것으로 기대된다.

ABSTRACT

The online portal platform provides online news with online comments, but the anonymity of comments causes incivility, and online comments are considered social problems. While there are many foreign language-based incivility detection studies, in-depth research is not being conducted in Korea since there has not been implemented Korean language dataset which is labeled detailed criteria of incivility. In this study, the incivility notation of comments was conducted in a total of 13 items, uncivil words were summarized. Furthermore, Attention algorithm was applied to each comment and summary to extract embedding vectors. 2-d CNN followed at the end to detect incivility in given data. As a result, we showed that the proposed algorithm is useful for anti-citizen detection such as name-calling and offensive tones. This study is expected to contribute to the formation of a healthy online comment culture by detecting uncivil comments which hinder democratic discourse.

키워드 : 온라인 댓글, 반시민성 검출, 어텐션 알고리즘, CNN

Keywords : Online commnets, detection of incivility, Attention Algorithm, CNN

Received 30 September 2022, **Revised** 6 October 2022, **Accepted** 13 October 2022

*** Corresponding Author** Hee-Jo Keum(E-mail: hkeum@skku.edu, Tel: +82-2-760-0689)
Professor, Department of media communication, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.12.1880 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

대부분의 정보가 디지털 콘텐츠화 되어 온라인 포털 플랫폼을 통해 유통되면서, 언론사들 역시 자체 홈페이지를 운영하면서도 온라인 포털 플랫폼으로 기사를 전달하고 있다. 온라인 포털에 실린 기사와 함께 형성되는 온라인 댓글은 사회적 여론을 만들어낼 뿐 아니라, 사회적으로 영향력을 가지게 되었다[1]. 온라인 댓글은 사회적인 논의가 필요한 문제들을 공론화시켜 적절한 사회적 조치가 이루어지게 하는 긍정적 효과도 가져왔지만 댓글의 익명성은 반대 의견에 대한 공격적 표현, 혐오적 표현 등의 반시민적인 표현까지도 허용하게 되어 현재는 또 다른 디지털 사회의 문제점으로 간주되고 있다[2].

이와 같은 사회적 문제점을 해결하기 위하여 포털 플랫폼은 자체적으로 비난이나 욕설과 같은 혐오적 표현을 감지하여 삭제하는 '클린봇' 기능을 운영하고 있다. 국내 최대 포털사이트인 네이버 역시 2020년 이후 인공지능형 클린봇을 운영하며 이전 버전보다 더 나은 자연어 처리 능력과 혐오 표현 검출율을 선보이고 있다[3]. 이로 인해 해당 포털의 댓글에서는 노골적인 비난이나 욕설 표현이 상당 부분 걸러진 채 게시되고 있다.

클린봇 등장 이후 댓글들은 필터링을 우회하기 위해 노골적 표현보다는 인신공격성 표현을 사용하는 무례한 표현, 상대의 생각을 근거없이 거짓이라 치부하는 표현, 느낌표(!)나 물음표(?)와 같은 문장 부호를 연속적으로 사용하여 나타내는 공격적 어조 등 반시민적 특성을 더 많이 나타내고 있다[4-6]. 반시민적인 댓글은 온라인 기사에 대한 관심은 높일 수 있지만 건전한 민주주의적 담론을 형성하는 데에는 부정적인 영향을 미치게 된다[7]. 무례하고도 공격적인 반시민적인 댓글은 미디어에 대한 대중의 신뢰도를 떨어뜨리게 되고 상대 의견에 대해 개방적인 마음(Open-minded)을 갖지 않게 될 뿐만 아니라 적대감이나 분노까지도 유발하는 것으로 연구되고 있다[8-12].

그러나 포털사이트의 인공지능형 클린봇 뿐 아니라, 머신러닝을 기반으로 하는 자연어 처리 연구들 역시 댓글의 반시민성보다는 단어 자체에 혐오 표현이 명시적으로 드러나 있는 비난, 욕설, 차별적 표현을 검출하는 데에 초점을 두고 있다[2, 13, 14]. 그렇기 때문에 공격적 어조나 특정 개인이나 집단에 대한 무례한 표현 등의 반시민적 표현은 최근까지도 댓글에서 나타나고 있으며 특히 정치기사 등의 댓글에서 더욱 빈번하게 관찰되고 있다.

온라인 댓글에서의 혐오 표현보다 반시민적 표현의 검출에 초점을 맞춘 연구가 등장하기 위해서는 먼저 댓글 표현들의 반시민성을 분류하기 위한 코딩 기준이 마련되어야 하고 이 분류에 따라 라벨링이 이루어진 댓글 데이터셋이 구축이 되어야 한다. 본 연구에서는 반시민적인 표현이 가장 빈번하게 나타나게 되는 정치 기사, 특히 20대 대통령 선거 캠페인 기간 동안 발행된 기사의 댓글 데이터를 수집하여 총 13가지의 세부 항목으로 이루어진 반시민성 분류 코딩 기준을 세우고 이에 따라 라벨링한 댓글 데이터를 구축하였다. 또한 댓글의 반시민성 탐지를 위한 어텐션 기반 분류 모델도 함께 제안함으로써 온라인 포털 플랫폼에 만연한 댓글의 반시민성을 경감시키고 건전한 온라인 담론 문화를 형성할 수 있는 기반을 마련할 것으로 보인다.

Ⅱ. 관련 연구

온라인 기사의 댓글 데이터에서 반시민성을 검출하기 위한 연구에서 가장 중점을 두는 부분은 데이터셋의 구축이다. 자연어로 이루어진 댓글 데이터는 분류 클래스 간의 애매모호성을 최소로 하면서 사전적 및 맥락적 의미가 댓글 간에서도 유지될 수 있도록 라벨링을 하는 작업이 반드시 필요하다. 클래스 간에 명확한 구분이 없이 서로 교집합을 이루는 부분이 있게 되면 뛰어난 분류 알고리즘을 적용하더라도 오검출율이 높게 나타나게 된다[15].

온라인 댓글의 반시민성을 세부 항목으로 분류한 후 각 항목마다 분류 기준을 세운 연구 [16]에서 사용하였던 댓글 데이터를 그대로 사용한 국외 연구[17]은 [16]의 세부항목 중 개인이나 집단에 대한 무례한 호칭(Name-Calling), 맥락 없이 음란하거나 천박한 단어를 사용한 표현(Vulgarity) 두 가지 항목을 검출하는 모델을 제시하였다. FastText를 사용하여 단어 벡터로 임베딩(Embedding) 된 댓글은 양방향 재귀모델 알고리즘과 Convolutional Neural Network(CNN)을 연속적으로 통과한 후 시그모이드 활성화 함수에서 이진 분류(Binary Classification)값으로 변환되며 무례한 호칭(Name-Calling)이나 천박한 욕설(Vulgarity) 항목 클래스로 분

류되었다. 그 결과, 무례한 호칭(Name-Calling)에 대해서는 f1-score값이 0.48이었으며 천박한 욕설(Vulgarity)에 대해서는 0.52로 나오게 되었다.

국외 연구[18] 또한 [16]의 연구와 유사하게 연구 [17]의 뉴스 기사 댓글 데이터셋을 사용하였다. 다만, 제안하는 알고리즘의 강인함을 보여주기 위하여 정치 트위터 댓글 데이터를 함께 사용하였다. 댓글 데이터는 BERT(Bidirectional Encoder Representations from Transformers)를 사용하여 워드 임베딩 되었다. BERT는 셀프 어텐션(Attention)을 기반으로 하는 인코더가 디코더에 연결된 트랜스포머 계열의 인코더-디코더 seq2seq 언어 모델로써 디코더가 출력을 만들어낼 때 인코더로 들어온 문장 전체를 참조하여 맥락을 고려하도록 설계되어 같은 단어라도 그 맥락에 따라 다른 벡터변수로 임베딩 된다는 강력한 장점이 있는 모델이다[19, 20]. 제안한 모델로 댓글의 반시민성을 분류한 결과, 연구[16]의 뉴스 데이터에서는 f1-score 0.65, 정치 트위터에서 0.80, 러시안 트롤 트윗 데이터에서는 0.81의 정확도를 보여주게 되었다.

이 외에도, 트위터의 댓글들을 TWEM(Transformed Word Embedding Model) 언어 모델을 사용하여 워드 임베딩한 국외 연구[21]은 성차별, 인종 차별, 공격적 표현, 혐오 표현을 분류하여 각각 f1-score 0.76, 0.78, 0.96, 0.49의 정확도를 보여주었다.

이 외에도 많은 국외 연구들은 댓글의 반시민성의 세부 항목 분류에 초점을 두고 다양한 딥러닝 모델을 활용하여 워드 임베딩을 진행하고 있다. 다만, 아직까지는 대부분의 연구들이 공격적 표현, 차별적 표현, 욕설과 같이 단어 자체가 반시민적 특성을 띄는 항목들을 위주로 분류하는 양상을 보이고 있다.

반면, 한글 데이터를 수집하여 반시민성을 상세 분류해내는 국내 연구는 국외에 비해 많지 않다. 특히 모아진 댓글 데이터에 대해 반시민적 표현에 대해 명확한 코딩 기준을 만들고 이에 따라 라벨링까지 모두 진행한 연구의 수는 손에 꼽힌다. 국내 연구 [14]은 9.4K에 달하는 온라인 연예 뉴스 댓글을 수집하여 혐오 표현, 차별적 표현, 공격적 표현을 수동으로 라벨링함으로써 한국말 혐오 표현 사전을 구축하기도 하였다. 이후 많은 국내 연구들이 이 데이터를 참조하여 혐오 표현을 분류해낼 수 있는 모델링을 진행할 수 있었다. 국내 연구[22]는 연구[14]에서 만든 한국어 혐오 표현 사전을 바탕으로 비난 및 공격적 표현(Hate, Offensive), 욕설(Hate, Abuse), 성차별(Gender), 정치적 차별 (Politics), 연령차별(Age), 종교차별 (Religion), 인종차별 (Race) 의 7가지 영역에 대해 이진 분류하는 모델을 만들었다. 이 모델에서는 CNN을 분류 알고리즘으로 사용하였는데, CNN의 출력 부분에 7가지 영역의 이진 분류기를 연결하고 입력 부분에는 세 종류의 특징추출벡터(워드 임베딩)가 들어오도록 구현하여 성능을 높이고자 하였다. 그 결과, 성차별(Gender)과 Politics(정치적 차별) 항목을 제외한 나머지 항목에서 Ko-BERT(한국어 BERT)나 LSTM을 적용했을 때보다 더 나은 성능을 보여주었다. 성능 평가는 f1-score로 측정되었는데, 욕설 표현에서 가장 높은 f1-score값인 0.7955 값이 나왔고 인종 차별에서 0.6098이 나왔다.

이 외에도 다양한 혐오 표현을 검출해내기 위해 SVM을 적용하여 의미 분석을 하거나 CNN을 적용한 연구들이 이루어지기도 하였고 한국어로 된 비윤리적 단어 사전의 구축에 대한 연구도 이루어지기 하였으나[23-25] 국내 연구에서는 본 연구에서 분류하고자 하는 상세 분류가 라벨링 되어 있는 데이터셋이 존재하지 않았다.

Ⅲ. 제안 기법

3.1. 어텐션(Attention) 기반 특징 추출 알고리즘과 멀티 채널 CNN

3.1.1. 어텐션 계층

어텐션 메커니즘은 순환 뉴럴 네트워크(Recurrent Neural Network)의 인코더 디코더 seq2seq의 구조를 갖춘 워드 임베딩 언어 모델이다. 다만, 디코더의 출력을 내보내는 과정에서 디코더의 타겟 데이터만 참조하는 것이 아니라 다시 인코더의 정보를 같이 참조하여 새로운 벡터를 만들어낸다는 점에서 그 차이점이 있다. 어텐션 메커니즘은 트랜스포머 알고리즘의 기반을 이루고 있으며 앞선 연구들에서 워드 임베딩 방법으로 쓰인 BERT 역시 트랜스포머 계열의 워드 임베딩이라고 할 수 있다[19, 20, 25].

어텐션 알고리즘은 인코더의 정보를 다시 참고하는 과정에서 어텐션 스코어라는 값을 구하는 과정이다. 어텐션 스코어는 현재 '디코더'를 통해 찾아낸 답과 인코더의 입력 문장 간의 유사도를 나타낸다. 초기 어텐션

메커니즘은 인코더의 은닉 상태(hidden state)와 디코더의 은닉 상태(hidden state)의 '내적'이라는 방법을 통해 유사도를 계산하였다.

$$Attention\ Score\ A = d_t^T e_n \quad (1)$$

어텐션 스코어는 인코더 내부의 셀의 개수만큼 나오게 되며 나온 값은 소프트맥스 함수에 각각 통과시켜 어텐션 가중치로 나오게 된다. 어텐션 가중치는 어텐션 분포 하에 존재하게 되는 것을 가정하므로 가중치의 총합은 1이 되도록 얻어진다. 특정 시점 t에서 최종적으로 얻어지는 어텐션 알고리즘은 n개의 인코더 셀에서 나오는 어텐션 가중치와 각 셀의 은닉 상태들을 곱하여 합한 형태로 얻어진다.

$$Attention\ \ a_t = \sum_{i=1}^{n} A_i^t h_i \quad (2)$$

3.1.2. 멀티 채널 Convolution Neural Network (CNN)

CNN은 분류 성능이 높아지는 방향으로 필터의 가중치 값을 학습시켜가는 알고리즘을 의미한다[26]. 본래는 이미지의 분류 알고리즘으로써 개발되었으나 최근 순환 알고리즘(RNN)과 함께 자연어 처리 분야에서도 많이 쓰이고 있다[27]. CNN은 두 개의 모듈로 구성되어 있는데, 하나는 필터를 합성곱하여 특징 벡터로 만들어내는 합성곱 계층(convolutional layer)이고 다른 하나는 만들어진 특징 벡터 중 대표 값을 설정하는 풀링 계층(pooling layer)이다. 풀링 방법으로는 최대값을 국소 영역의 대표값으로 쓰는 Max Pooling 이나 평균값을 사용하는 Average Pooling등이 있다.

컬러 이미지의 경우 보통 세 개의 채널(R, G, B)로 이루어져 있기 때문에 CNN의 합성곱 계층은 채널 별로 필터를 적용하는 멀티 채널 구조를 가지게 된다. 각 채널은 독립적으로 입력 이미지와 필터의 합성곱 결과를 도출해 내고, 학습이 진행될 때 각 필터의 가중치는 채널별로 다르게 나온다. 연구 [27, 28]은 이러한 CNN의 특성을 이용하여 자연어 처리를 진행하였다. [27]의 경우, 어텐션 메커니즘을 Bi-LSTM의 은닉 상태에 적용함으로써 CNN의 멀티 채널 입력을 만들었다. [28]은 입력 데이터의 워드 임베딩을 다섯 가지로 다르게 한 후 각 채널의 입력으로 사용하였다. 그 결과, SVM을 이용한 분류한 것보다 항목별로 2.7%, 5.6% 정도 f1-score얻을 수 있었다.

본 연구에서는 멀티 채널 CNN의 특징을 사용하여, 두 개의 채널이 다른 입력 데이터를 받도록 하는 구조를 제안하고 있으며, 각각의 입력은 문장 데이터와 해당 문장의 요약 데이터로 지정하였다. 각각의 채널로 들어오는 문장 데이터 및 요약 데이터는 어텐션 메커니즘을 적용하여 임베딩 되었으며 이를 기반으로 중복을 허용하는 13개로 라벨 분류가 진행되도록 하였다[5].

기존 연구에서 멀티 채널 CNN은 n-gram 워드 임베딩을 각각의 채널로 적용하는 데에 사용되었다[27]. 그러나 n-gram은 문장 내에서 n개의 단어만 고려하기 때문에 문장 전체의 맥락을 반영하기가 어렵고 n개의 단어 순서 그대로 다른 문장에서 나올 확률도 낮기 때문에 수많은 사용자가 문법이나 어법을 지키지 않고 작성하는 댓글 분류 연구에는 적합하지 않다. 본 연구에서는 원문 데이터에 어텐션 메커니즘을 적용함으로써 문장 전체의 맥락에서의 포지션 정보가 반영되도록 하면서 요약 데이터를 사용하여 문장 판별의 기준이 되는 부분을 함께 제공함으로써 기존 연구의 한계점을 극복하려고 하였다.

Ⅳ. 실 험

4.1. 데이터의 수집 및 전처리

댓글의 반시민적 표현 분류를 위한 모델링을 진행하기 위해, 댓글들 중 가장 반시민적 표현이 많이 나타나는 정치 기사의 댓글을 타겟으로 하였다. 특히 20대 대통령 선거 기간 동안 발행된 뉴스 기사는 지지층 사이에서 발생하는 댓글들의 반시민적 표현의 정도가 매우 심각하였기 때문에 본 연구에서는 해당 기사들의 제목, 내용 및 맥락 정보들과 댓글 정보를 함께 수집하게 되었다. 총 기사의 수는 중복을 제외하고 18,393개였으며 이 중 핸드 코딩을 위해 18,393개의 기사 중 기사 길이 2000자 이내, 댓글 수 100개 이상인 것 200개를 무작위로 선정하게 되었다.

수집된 댓글들은 문장 단위로 나뉘어 졌으며 나뉘어진 문장들은 토크나이저의 단어 사전을 먼저 학습시켜 신조어나 줄임말 등이 형태소로 분해되지 않도록 하였다. 본 연구에서는 Python 라이브러리인 kiwipiepy 토크나이저를 사용하여 토큰화 한 후 고유명사/대명사/형용사/부사/느낌표 및 물음표가 나타난 어미/등으로 품사가 태깅된 단어들 중 불용어 리스트에 없는 단어들만 데이

터에 남아있도록 하였다.

4.2. 반시민성 코딩 기준과 텍스트 요약 및 라벨링

Table. 1 Coding Scheme of Incivility

Labeling Category		Description
offensive & rude expression	name-calling	a derogatory expression or derogatory expression for a particular individual and a particular group
	lying	Denying the integrity of an individual or group immediately or stating that a policy or idea is groundless false.
	offensive violent imperative tone	an oppressive tone screaming expression
	Vulgarity	coarse words
hate speech	speech which attacks Democratic discourse	discriminatory remarks against the region
		discriminatory remarks against the disables
		discriminatory remarks against the gender
		discriminatory remarks against the race and immigrants
		discriminatory remarks against the religion
		discriminatory remarks against the LGBTQs
		discriminatory remarks against the occupation
		discriminatory remarks against the age
		discriminatory remarks against the supporters of a particular political party

4.2.1. 반시민적 표현 코딩 기준

반시민성 코딩 기준은 연구 5]를 토대로 하여 세워졌으며 이에 추가하여 차별적 발언의 영역으로 지역, 성별, 지지하는 정당, 연령 등에 대한 차별 발언을 포함시켜 더욱 세밀하게 분류되도록 하였다.

코딩 기준은 표1에 대략적으로 정리되어 있으나, 실질적으로 코딩을 진행할 때에는 각 항목 간 의미적 거리가 가장 멀어지도록 세부 기준을 더 세워 애매모호성을 최소화 하였다.

4.2.2. 텍스트 요약 및 라벨링

Table. 2 Incivility frequencies

Labeling Category		Numbers
offensive & rude expression	name-calling	8941
	lying	976
	offensive violent imperative tone	3276
	Vulgarity	286
hate speech	against the region	141
	gainst the disables	90
	against the gender	251
	against the race and immigrants	57
	against the religion	36
	against the LGBTQs	16
	against the occupation	436
	against the age	369
	against the supporters of a particular political party	647

본 연구에서는 라벨링을 진행하면서 문장 내에서 반시민적인 표현을 이루고 있는 단어나 구를 '요약' 데이터로서 표기하도록 하였다. 예를 들어, '공격적 표현' 에 해당한다고 표기하였다면 어떤 단어에서 '공격적 표현' 으로 여겨졌는지를 표기하는 것이다. 만일 하나의 문장 안에 여러 가지의 표현이 존재한다면 순서대로 모두 표기하도록 하여 원문 데이터에서 나타난 단어의 순서와 동일하도록 표기하였다. 원문 데이터와 요약 데이터는 각각 멀티채널 CNN의 입력으로 사용되었다.

라벨링이 표1의 코딩 기준대로 이루어지도록 5명의 코더가 한달 정도의 훈련을 받았으며, 이후 커뮤니케이션 전공의 전문가들의 리뷰로 다시 한 번 데이터 라벨링과 요약 데이터를 검증하였다. 코더들은 한 문장에 대해 여러 개의 반시민성이 나타났을 경우 모두 표기하였으며 총 28,796개의 문장 중 반시민적 표현을 가진 문장은 총 11,923개였으며 표2에서 11,923개의 문장에서 나타난 반시민적 표현의 세부 항목의 빈도수가 나타나 있다. 개인과 집단에 대한 폄하나 비방, 무례한 표현의 경우 8,941건으로 가장 많은 수를 차지하고 있으며 공격적 어조가 그 다음으로 많은 수를 가지고 있다. 반면, 종교에 대한 차별 발언이나 성소수자에 대한 차별 발언은 각각 36건/ 16건으로 적은 편으로 나타났다. 이와 같은 수치 차이는 [22]에서도 나타났듯, 하나의 분야에만 해당하는 기사의 댓글들만을 수집하였기 때문으로 보인다. 라

Table. 3 Examples of coded uncivil comment and summary

example	offensive & rude				hate speech									summary
	name-calling	lying	offensive speech	Vulgarity	region	disables	gender	race	religion	LGBTQs	occupation	age	political party	
동족상잔 비극 만든 중공 짱깨들 삼부자한테는 평화공존 통일외치면서??????	0	0	1	0	0	0	0	1	0	0	0	0	1	중공, 짱깨들~서??????
지금 이 정권 보다 몇곱절 훌륭한 업적 남기셨고 그 좌파들의 선동에 지금까지의 고초를 겪으셨네요	0	1	0	0	0	0	0	0	0	0	0	0	1	좌파, 선동

벨링시 코더들은 각 항목에 대해 라벨링을 진행하면서 해당 세부 항목이 존재한다고 생각하게 된 단어나 구를 요약 데이터로서 표기하게 하였다. 코더들은 표 3과 같이 하나의 문장에 대해 여러 개의 세부 항목을 중복적으로 표기하였으며 마지막 열에 요약 데이터로서 반시민적인 표현을 적었다. 예를 들어, 표3에서 두 번째 문장의 경우 lying에 표기를 한 것은, 해당 문장 자체의 참 거짓을 판별하는 것이 아니라 다른 사람의 정책이나 말을 거

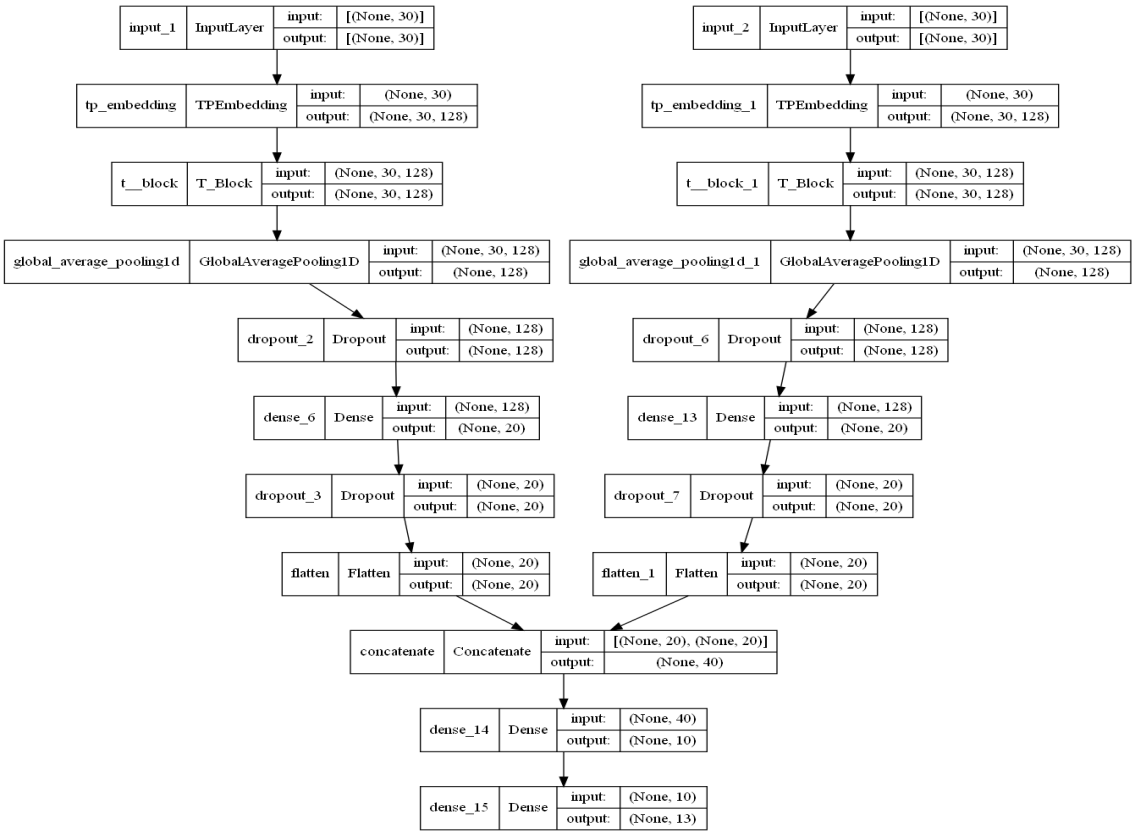

Fig. 1 Algorithm diagram of suggested method

짓 의도라고 폄훼할 때 표기하였으므로 요약에 '선동'이라고 적어둔 것이다. 댓글의 반시민적 특성은 한글로 표현되어야 하므로 표3에서는 한글로 예시와 요약을 표현하였다.

4.3. 어텐션 기반 멀티 채널 CNN

원문 데이터와 요약 데이터를 어텐션 계층에 통과시켜 얻은 2가지의 특성 채널로 그림 1과 같은 멀티채널 CNN을 구성하였다. Input_1과 Input_2는 각각 원문 데이터와 요약 데이터를 의미하며 token_and_positioning embedding 계층 및 transformer block 계층이 어텐션 계층을 의미한다. 각각의 input 은 동일하게 어텐션 계층을 통과하며, 이후 Conv1D(1D Convolution)과 max_pooling1D(1D maxpooling) 연산을 적용하고 나온 값을 연결하게 된다. 어텐션 알고리즘과 멀티 채널 CNN의 모든 구현은 파이썬의 텐서플로우 라이브러리를 사용하였으며 GPU를 사용하지 않고 훈련시켰다.

Ⅴ. 실험 결과

표4는 13가지의 반시민적 표현에 대해 학습된 결과를 요약하여 나타낸 것이다. 표에 나타난 Model-1과 Model-2는 원문데이터만이 하나의 입력으로 사용된 것이다. Model-1은 원문 데이터에 Embedding만 사용하여 1d CNN으로 분류한 것이고 Model-2는 원문 데이터를 어텐션 계층에 통과시켜 얻어진 특징 벡터를 1d CNN으로 분류한 것이다. Model-3은 원문 데이터와 요약 데이터 모두를 각각 어텐션 계층에 통과시켜 사용한 것이다(활성화함수: relu). 각 모델은 무례, 정책폄훼, 공격적 어조 등 총 13가지의 항목에 대해 멀티라벨 분류를 시행하였으며 각각의 항목에 대한 f1-score 및 전체의 f1-score는 테스트에 사용된 샘플의 수를 고려하여 계산되었다(CNN의 활성화 함수 : softmax, epoch=30). 본 데이터는 각 반시민성 항목들에서 반시민성에 해당하는 데이터의 수와 그렇지 않은 데이터의 수의 불균형이 심하기 때문에 단순한 정확도를 비교하면 제대로 결과를 파악할 수가 없으므로 f1-score를 지표로 정하였다. 테스트에 쓰인 댓글 문장의 수는 전체 댓글 수의 20%가 되도록 데이터를 나누었며 각 항목에 사용된 테스트 샘플의 수는 N열에 나타나 있다. 각 항목에 사용된 테스트 샘플의 수는 N열에 나타나 있다.

Table. 4 f1-score comparation : Model-1 uses original comments and applied 1-d CNN only. Model-2 uses original comments and applied Attention and 1-dCNN. Model-3 uses original comments and summary, applied Attention and 2-d CNN

item	Model-1	Model-2	Model-3	N
name-calling	0.85	0.81	0.86	1761
lying	0.34	0.18	0.35	193
offensive violent imperative tone	0.47	0.45	0.71	635
Vulgarity	0	0.03	0.08	61
against the region	0.44	0.26	0.5	25
gainst the disables	0.33	0	0.33	14
against the gender	0.41	0.41	0.35	54
against the race and immigrants	0	0.12	0.13	14
against the religion	0	0	0	5
against the LGBTQs	0	0.67	0.67	2
against the occupation	0.48	0.31	0.29	91
against the age	0.51	0.18	0.41	70
against the supporters of a particular political party	0.44	0.27	0.24	146
Average	0.68	0.65	0.76	3071

표4에서, 제안하는 Model-3의 평균 f1-score는 0.76으로써 다른 두 개의 모델의 f1-score(Model-1: 0.68, Model-2: 0.65)에 비해 더 높은 분류 성능을 보여주었다. 평균 f1-score는 각 항목에 해당하는 테스트 샘플의 수를 고려하여 구해졌기 때문에 테스트 샘플 중 가장 수가 많은 무례 항목에서의 f1-score가 전체의 평균에 영향을 많이 미치게 되어 있다. 그래서 Model-1의 경우 불필요한 욕설, 인종차별, 소수자 차별, 종교 차별의 항목에서 f1-score가 0인데도 불구하고 무례 항목에서의 f1-score가 높기 때문에 Model-2보다 높은 평균 값이 얻어졌다.

세부 항목 중 공격적 어조는 특정 단어로는 구별하기가 어렵다. 예를 들어,'죽고 싶냐'라는 댓글의 경우 차별적 단어나 무례한 호칭, 욕설 등이 들어있지 않으나 민주적인 담론을 저해하는 공격적 어조를 가진 반시민성으로 평가된다. 그러나 특정 단어가 들어 있지 않기 때문에 Model-1이나 Model-2에서는 각각 0.47, 0.45의 높

지 않은 f1-score를 보여주게 되었다. 그러나 Model-3에서는 0.71의 f1-score로서, 다른 두 모델에 비해 월등히 높은 성능을 보여주었다. 이는 요약 데이터의 사용이 모델의 성능 향상에 영향을 미치고 있다는 것을 나타낸다.

불필요한 욕설과 인종 차별 및 종교차별, 소수자 차별의 경우 세 개의 모델이 비슷한 성능을 보이고 있다. 이는 훈련에 이용된 데이터의 수가 다른 세부 항목에 비해 현저히 적었기 때문이다. 특히 불필요한 욕설의 경우, 포털에서 운영하는 클린봇과 같은 자체 필터가 미리 걸러내는 경우가 많기 때문에 라벨링 기준에 해당하는 수준의 불필요한 욕설은 거의 존재하지 않았다. 또한 정치 기사의 댓글이기 때문에 인종차별이나 종교차별, 소수자 차별의 발언보다는 정치인이나 반대편 입장의 그룹을 무례한 호칭으로 부르거나 그들의 정책이나 입장에 대해 거짓이라 폄하하는 댓글들이 주를 이루고 있다. 그러나 Model-3의 경우 인종차별이나 소수자차별의 항목의 N수가 각각 전체 데이터의 0.004, 0.0006의 비율에 해당함에도 f1-score가 0보다 큰 값을 나타내고 있기 때문에 추가로 데이터를 구성할 경우 이보다 더 높은 수준의 성능을 낼 수 있을 것이라고 예측된다.

Ⅵ. 결 론

본 연구에서는 20대 대선 관련 온라인 뉴스의 댓글 데이터를 바탕으로 반시민적 표현 탐지를 위한 모델을 개발하였다. 기존의 혐오 표현이나 욕설 등을 검출해내는 연구들은 심한 욕설이나 명시적인 혐오 표현을 검출해내는 것에는 잘 동작하였으나 맥락적 고려가 필요한 공격적 어조나 상대의 아이디어에 대한 맹목적 폄하, 개인에 대한 무례한 표현 등은 구별해내지 못하였다. 본 연구에서는 반시민적인 표현 13가지를 구별해내기 위하여 반시민성 코딩 기준을 세운 후 총 28796개의 댓글 데이터를 구축하여 각 댓글의 반시민성을 평가하였다. 또한 반시민성 표현에 대한 검출 성능을 높이기 위해 기존 CNN 구조에 원문 데이터와 요약 데이터에 어텐션 계층을 결합한 특성 채널을 각각의 입력으로 하는 모델을 제안하였다. 그 결과, 원문 데이터만을 이용한 모델에 비해 평균적으로 더 높은 성능을 확인할 수 있었다.

본 연구에서 제안하는 모델을 사용한다면 온라인 상에서의 혐오 및 차별 발언 뿐 아니라 민주주의적 담론 형성을 해치는 공격적 어조, 서로의 아이디어를 이유없이 비난하거나 폄하하는 표현, 개인을 무례하게 부르는 표현 등도 탐지하여 제한할 수 있다. 또한 일괄적 통제나 삭제의 방법 보다는 사용자에게 '반시민성 표현'에 대한 설명과 함께 다른 표현을 제안할 수도 있어 좀 더 민주적인 방법으로 온라인 댓글 문화를 형성할 수 있을 것이라 기대한다.

본 연구에서 제안한 반시민성 검출 모델은 다음과 같은 한계점을 가지고 있다. 첫째로, 13가지 항목 간 데이터 불균형이 있다는 점이다. 대선 관련 기사와 댓글 데이터를 활용하였기 때문에 상대의 후보나 정치인을 무례하게 표현하는 댓글이 대부분을 차지하였고 그 외의 표현은 상대적으로 매우 적은 수를 가지고 있어서, 인종차별, 종교 차별, 성소수자 차별 등에 대한 댓글에 대해서는 학습이 충분하게 이루어지지 못했을 수 있다. 텍스트 분류에서는 말뭉치의 양에 따라 모델의 성능이 달라질 수 있으므로 향후 연구에서는 본 연구에서 쓰인 샘플 데이터가 아니라 18393건의 전체 데이터에 대하여 데이터를 구축하여 모델의 성능을 개선시켜볼 예정이다.

두 번째 한계점은 요약 데이터의 작성 과정이 필요하다는 점이다. 어텐션 매커니즘은 워드 데이터의 맥락 정보를 제공해주는 강점이 있으나 Model-2의 결과에서 나타나 있듯이, '공격적 어조'와 같이 특정 단어의 유무 보다는 문장을 이루는 여러 단어들의 조합으로 반시민성을 나타내는 항목에서는 좋은 성능을 내지 못한다. 반면 Model-2에 요약 데이터의 어텐션 임베딩 벡터를 추가하여 원문의 어텐션 임베딩 벡터와 연결, 하나의 특징 벡터로 만든 Model-3은 반시민성 항목으로 평가하게 된 핵심 단어들의 맥락 정보까지도 분류에 이용하는 것이므로 더 높은 성능을 낼 수 있게 되었다.

본 연구에서는 200개 기사에서 나온 댓글 데이터로만 먼저 데이터셋을 구축하였지만 모델의 성능 개선을 향상시키기 위해서는 더 큰 데이터셋이 필요하다. 그러나 큰 데이터 셋의 요약 데이터를 만드는 과정은 코더의 훈련 과정이나 시간 면에서 쉬운 일이 아니다. 그렇기 때문에 추후 연구에서는 어텐션 메커니즘을 활용하여 데이터셋의 요약 데이터를 먼저 추출해내는 프로세스를 추가 적용할 것이다. 또한 현재 알고리즘이 차별적 발언에서는 좋은 결과를 내지 못하므로 차별적 표현을 가진 말뭉치 데이터를 라벨링하여 적용할 필요가 있다. 그렇게 된다면 본 모델과 데이터셋을 활용하여 건전한

온라인 댓글 문화 형성에 기여할 수 있는 검출 알고리즘 및 챗봇까지도 제작할 수 있을 것이라 기대된다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021S1A5C2A02088387

References

[1] Hankook Research. Toxic Comments, is it okay? [Internet]. Available: https://hrcopinion.co.kr/archives/14589.

[2] K. Kenski, K. Coe, and S. A. Rains, "Perceptions of Uncivil Discourse Online: An Examination of Types and Predictors," *Communication Research*, vol. 47, no. 6, pp. 795-814, Apr. 2020.

[3] S. H. Lee, "Biased Artificial Intelligence: Analyzing the Types of Hate Speech Classified by 'Cleanbot', NAVER AI for Detecting Malicious Comments," *Journal of Cybercommunication Academic Society*, vol. 38, no. 4, pp. 33-75, Dec. 2021.

[4] P. Rossini, "Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk," *Communication Research*, vol. 49 no. 3, pp. 399-425, May 2022.

[5] K. Coe, K. Kenski, and S. A. Rains, "Online and uncivil? Patterns and determinants of incivility in newspaper website comments," *Journal of Communication*, vol. 64, no. 4, pp. 658-679, Jun. 2014.

[6] Z. Papacharissi, "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups," *New media and society*, vol. 6, no. 2, pp. 259-283, Apr. 2004.

[7] S. Wright and J. Street, "Democracy, deliberation and design: the case of online discussion forums," *New media and society*, vol. 9, no. 5, pp. 849-869, Oct. 2007.

[8] A. A. Anderson, D. Brossard, D. A. Scheufele, M. A. Xenos and P. Ladwig, "The nasty effect: Online incivility and risk perceptions of emerging technologies," *Journal of computer-mediated communication*, vol. 19, no. 3, pp. 373-387, Apr. 2014.

[9] P. Borah, "Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere," *Communication Research*, vol. 41, no. 6, pp. 809-827, Aug. 2014.

[10] S. Agarwal and A. Sureka, "A focused crawler for mining hate and extremism promoting videos on YouTube," *In Proceedings of the 25th ACM conference on Hypertext and social media*, pp. 294-296, Sep. 2014.

[11] B. T. Gervais, "Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment," *Journal of Information Technology and Politics*, vol. 12, no. 2, pp. 167-185, Jan. 2015.

[12] G. M. Masullo and J. Kim, "Exploring "angry" and "like" reactions on uncivil Facebook comments that correct misinformation in the news," *Digital Journalism*, vol. 9, no. 8, pp. 1103-1122, Oct. 2021.

[13] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: a survey on multilingual corpus," *In 6th International Conference on Computer Science and Information Technology*, vol. 10, pp. 10-5121, Feb. 2019.

[14] J. H. Moon, W. I. Cho, and J. B. Lee, "Beep! Korean Corpus of Online News Comments for Toxic Speech Detection," *In Proceeding of the 8th International Workshop on Natural Language Processing for Social Media*, Taipei, 2020.

[15] A. Stoll, M. Ziegele and O. Quiring, "Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments," *Computational Communication Research*, vol. 2, no. 1, pp. 109-134, Feb. 2020.

[16] K. Coe, K. Kenski and S. A. Rains, "Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments," *Journal of Communication*, vol. 64, no. 4, pp. 658-679, Jun. 2014.

[17] F. Sadeque, S. Rains, Y. Shmargad, K. Kenski, K. Coe and S. Bethard, "Incivility detection in online comments," in *Proceedings of the eighth joint conference on lexical and computational semantics*, pp. 283-291, 2019.

[18] K. B. Ozler, K. Kenski, S. Rains, Y. Shmargad, K. Coe, and S. Bethard, "Fine-tuning for multi-domain and multi-label uncivil language detection," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Online, pp. 28-33, 2020.

[19] W. Liu, L. Li, Z. Huang, and Y. Liu, "Multi-lingual Wikipedia Summarization and Title Generation on Low Resource Corpus," in *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, Varna, Bulgaria, pp. 17-25, 2019.

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert:

Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, Minneapolis: MN, USA, 2019.

[21] R. Kshirsagar, T. Cukuvac, K. McKeown, and S. McGregor, "Predictive Embeddings for Hate Speech Detection on Twitter," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, pp. 26-32, 2018.

[22] W. Lee and H. Lee, "Bias & Hate Speech Detection Using Deep Learning: Multi-channel CNN Modeling with Attention," *Journal of the Korea Institute Of Information and Communication Engineering*, vol. 24, no. 12, pp. 1595-1603, Dec. 2020.

[23] J. Hong, S. Kim, J. Park, and J. Choi, "A Malicious Comments Detection Technique on the Internet using Sentiment Analysis and SVM," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 20, no. 2, pp. 260-267, Feb. 2016.

[24] Y. Kim, H. Kang, S. Han, and H. Jeong, "Swear Word Detection through Convolutional Neural Network," in *Proceedings of the Annual Spring Conference of KIPS*, vol. 28, no. 2, pp. 685-686, 2021.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv preprint arXiv:1409.0473*, 2014.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[27] Z. Liu, H. Huang, C. Lu, and S. Lyu, "Multichannel CNN with Attention for Text Classification," *arXiv preprint arXiv:2006.16174*, 2020.

[28] C. Quan, L. Hua, X. Sun, and W. Bai, "Multichannel Convolutional Neural Network for Biological Relation Extraction," *BioMed Research International*, vol. 2016, Article ID. 1850404, Dec. 2016.

박윤정(Youn-Jung Park)

연세대학교 인지과학협동과정 공학박사
성균관대학교 글로벌 융복합 콘텐츠 연구소 선임 연구원
※관심분야 : 머신러닝, HCI

이세영(Se-Young Lee)

성균관대학교 미디어커뮤니케이션학과 교수
※관심분야 : 인간 소통 행동, 인간 챗봇 상호작용, 사회적 영향

금희조(Hee-Jo Keum)

성균관대학교 미디어커뮤니케이션학과 교수
※관심분야 : 방송, 영상, 미디어 효과