ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 45, No. 11, pp. 1185-1192, 2018. 11
https://doi.org/10.5626/JOK.2018.45.11.1185

퍼지 범주 표현과 준지도 심층 신경망을 이용한 트위터 혐오 발언 문장 탐지

(Semi-Supervised Learning for Detecting of Abusive Sentence on Twitter using Deep Neural Network with Fuzzy Category Representation)

박 다 솔 †　　차 정 원 ††
(Da-Sol Park)　　(Jeong-Won Cha)

요 약 사회관계망 서비스(SNS, Social Network Service)에서 발생하는 혐오 발언 문장으로 인해 피해를 보는 사람이 점점 늘고 있다. 본 논문은 트위터 문장에서 단순 사전 비교를 통한 혐오 발언 탐지를 넘어 문장의 내포된 의미가 혐오성인지 아닌지를 판단하기 위해 대용량의 파일에서 준지도 학습과 심층 신경망을 이용한 탐지 방법을 제안한다. 대부분 혐오 단어로 구성된 블랙리스트를 생성하여 이것과 비교하여 판단한다. 하지만 이러한 방법은 혐오 발언의 미묘하고 교묘한 표현을 찾아내지 못한다는 단점이 존재한다. 그리고 한국어 트위터 문장에 대해 혐오 발언 여부에 대한 레이블을 부착한 코퍼스를 생성하였다. 트위터 코퍼스 4만4천문장을 학습하였고, 1만3천여문장을 평가하여 음절 1-layer CNN과 문장 벡터를 사용한 모델의 결과가 명시적 혐오 발언의 F1 Score 86.13% 성능을 보였다. 음절 1-layer CNN과 2-layer CNN 그리고 문장 벡터를 사용한 모델 결과가 암시적 혐오 발언의 F1 Score 25.53%의 성능을 얻었다. 논문에서 제안한 방법을 이용하여 사이버 불링을 탐지하기 위한 방법으로 사용할 수 있다.
키워드: 혐오 발언, 퍼지 범주 표현, 준지도 학습, 자연어 처리, 기계 학습

Abstract The number of people embracing damage caused by hate speech on the SNS(Social Network Service) is increasing rapidly. In this paper, we propose a detection method using Semi-supervised learning and Deep Neural Network from a large file to determine whether implied meaning of sentence beyond hate speech detection through comparison with a simple dictionary in twitter sentence is abusive or not. Most of the methods judge the hate speech sentence by comparing with a blacklist comprising of hate speech words. However, the reported methods have a disadvantage that skillful and subtle expression of hate speech cannot be identified. So, we created a corpus with a label on whether or not to hate speech on Korean twitter sentence. The training corpus in twitter comprised of 44,000 sentences and the test corpus comprised of 13,082 sentences. The system performance about the explicit abusive sentences of the F1 score was 86.13% on the model using 1-layer syllable CNN and sequence vector. And the system performance about the implicit abusive sentences of the F1 score 25.53% on the model using 1-layer syllable CNN and 2-layer syllable CNN and sequence vector. The proposed method can be used as a method for detecting cyber-bullying.
Keywords: hate-speech, fuzzy category representation, semi-supervised learning, natural language processing, machine learning

· 이 논문은 2017~2018년도 창원대학교 자율연구과제 연구비 지원으로 수행된 연구결과임
· 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2018-0-00247, GANs를 이용한 딥러닝용 학습데이터 자가 증식 기술 및 유효성 검증 기술 개발)

† 학생회원 : 창원대학교 친환경해양플랜트FEED공학과
dasol_p@changwon.ac.kr
†† 종신회원 : 창원대학교 컴퓨터공학과 교수(Changwon Nat'l Univ.)
jcha@changwon.ac.kr
(Corresponding author임)

논문접수 : 2018년 7월 18일
(Received 18 July 2018)
논문수정 : 2018년 9월 3일
(Revised 3 September 2018)
심사완료 : 2018년 9월 7일
(Accepted 7 September 2018)

1. 서 론

청소년의 따돌림 문제가 심각한 사회문제로 대두되는 최근 스마트폰 보급과 인터넷 사용이 일상화되면서 학교폭력은 사이버상의 따돌림과 괴롭힘으로까지 확대되고 있으며 사이버 불링의 발생률이 증가하고 있다.

'사이버 불링'은 사이버 범죄의 다른 말로, 인터넷과 관련된 기술상에서 다른 사람에게 피해를 입히는 모든 행위를 의미한다[1]. 사이버 불링의 피해자는 자살을 선택할 정도로 심각한 물리적, 정신적 피해를 받는다.

그 중 혐오표현(Hate speech)의 정의는 인종, 성, 연령, 국적, 성 정체성, 장애, 언어능력, 도덕관 또는 정치적 견해, 사회적 계급, 직업 및 외모, 지적능력, 혈액형 등 특정한 그룹에 대한 편견, 폭력을 부추길 목적으로 이루어지는 의도적인 폄하, 위협, 선동 등을 담은 발언을 뜻한다[2]. 학자마다 혐오표현의 구체적인 정의에 대하여는 차이를 보이고 있다.

본 논문에서는 상대방에 대한 비방이나 욕설, 따돌림, 괴롭힘 등과 같은 언어 폭력적인 혐오표현이 포함된 발언을 '혐오 발언'이라고 정의한다. 이러한 발언이 실제 행동과 연관이 있기 때문이며 혐오 발언으로 인한 혐오 범죄가 증가하고 있기 때문에 혐오 발언 인식의 중요하다. 그리고 트위터(Twitter), 페이스북(Facebook), 유튜브(Youtube)와 같은 소셜 미디어 회사들은 혐오 발언을 찾기 위해 노력하고 있다.

혐오 발언을 탐지하기 위해서는 혐오 단어로 구성된 블랙리스트 매칭으로는 한계가 존재한다. 혐오 발언에 대한 정의가 불명확하며 혐오 발언의 영향에 대해 분석이나 조사가 이루어지지 않고 있으며 자연어 처리 연구에서 분류기 학습을 위한 학습 코퍼스가 존재하지 않는다.

또한 명확한 단어가 나타나는 혐오 표현이 있는 반면 암시적인 혐오 표현이 있다. 이것은 블랙리스트를 사용하여 탐지할 수 없다.

표 1과 같이 혐오 단어 매칭만으로 탐지하기 어려운 문장 즉, 혐오 단어가 포함되지 않은 혐오 문장을 탐지하는 것을 본 논문에서의 목표로 한다. 이러한 문장은 문장의 의미가 혐오성이 내포되어 있다는 것을 의미하기 때문에 암시적 혐오 문장(Implicit abusive sentence)라고 명칭한다. 다양한 유형이 나타날 수 있는 비정형 데이터를 적용하여 블랙리스트에 존재하지 않는 혐오 발언을 탐지하는 모델을 만들고, 이를 이용하여 사이버 불링 탐지가 가능하도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 소개하고, 3장에서는 제안 방법에 대해 상세히 기술한다. 4장에서는 실험을 기술하고 성능을 분석한다. 5장에서는 결론을 기술한다.

표 1 암시적 혐오 문장 예제
Table 1 Examples of implicit abusive sentences

Example
Do you have any money to be appointing a lawyer if you are sude? I guess that you no brain, you have nothing in hand.
Can you let the long needle pass through your neck?

2. 관련 연구

트위터의 일반 지침 및 정책 내 혐오 행위 관련 정책이 있으며 트위터 운영 원칙에 설명된 바와 같은 행위는 허용되지 않는다. 트위터 정책 내 명시된 혐오 행위는 인종, 민족, 국적, 종교, 성적 지향, 성별, 성 정체성, 종교, 나이, 장애, 질병 등을 이유로 타인에게 폭력적인 행위를 하거나, 직접적인 공격을 가하거나, 위협하는 발언 등을 포함한다. 표 2는 트위터에서 정의한 개인 또는 단체를 괴롭히는 행위이다.

[3]은 단어 임베딩과 문자 n-gram의 벡터를 concat하여 자질 임베딩을 생성하고 자질 임베딩을 입력으로 하는 CNN을 이용하여 혐오 발언을 분류하는 방법을 제안했다.

[4]는 BOW모델을 기반으로 한 paragraph2vec을 이용하여 혐오 발언을 탐지하는 논문이다. 새로운 단어에 대한 해결법은 이미 학습한 단어 임베딩을 삽입하는 방식을 이용하였다.

[5]는 다양한 자질과 탐지 기법을 사용한 지도학습과 비지도 학습을 결합하는 방식을 적용한 연구를 진행하였다.

[6]은 혐오 단어 학습기와 LSTM 분류기의 부트스트랩을 적용하여 혐오 발언을 탐지하는 실험을 진행했다. 혐오 단어 학습기는 설정 threshold보다 점수가 높으면 새로운 혐오단어로 인식되는 방식을 추가하는 방식을 적용하여 자동으로 혐오 단어를 늘리는 방법을 채택하여 연구를 진행하였다.

본 논문에서는 트위터 문장과 같은 비정형 텍스트에는 더욱 다양한 유형이 나타난다는 점을 토대로 하여 하나의 문장을 음절 단위로 형태를 변경하여 입력으로

표 2 개인 또는 단체를 괴롭히는 행위의 예시
Table 2 Examples of acts of bullying by individuals or organizations

- Intimidation
- The behavior of physical injury, death or illness of an individual or group
- The behavior method doing concrete harm violence or mass murder about the main target and victim group
- The behavior encourage fear about private group
- Repetitive or unilateral slander, abuse, racial or gender discriminatory utterance, or other offensive content

하는 CNN(Convolution Neural Network)을 이용하고 추가적으로 문장에 대한 자질을 사용하기 위해서 문장 벡터를 생성하여 혼합하여 사용하는 네트워크로 구성된다.

소량의 레이블이 부착되어 있는 데이터와 레이블이 부착되어 있지 않은 대량의 데이터를 이용하는 준지도 학습(Semi-supervised Learning) 또한 적용하였다.

3. 제안 방법

3.1 제안 방법

제안하는 방법은 크게 3가지로 구분되며 (1)초기 학습 코퍼스 구축, (2)CNN 학습, (3)세대(Generation) 생성이다. 그림 1은 본 논문에서 제안하는 방법을 요약하여 보여준다. 세부 설명은 3.2부터 3.5에 걸쳐서 자세하게 설명한다.

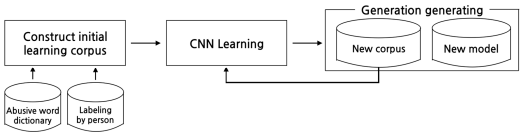

그림 1 제안 방법
Fig. 1 Proposed method

3.2 초기 학습 코퍼스 구축

학습을 진행하기 위해서는 학습 코퍼스에 정답 레이블이 부착되어 있는 코퍼스가 필요하다. 그러나 정답 코퍼스를 제작하는 것은 사람이 직접 생성해야하기 때문에 많은 시간과 비용이 소모된다. 이러한 점을 보완하기 위해서 반자동으로 정답 코퍼스를 생성하기 위한 방법이 필요하다.

혐오 단어 사전을 이용하여 각 트위터 문장에 대해 혐오 발언 여부에 대한 레이블을 부착하는 작업을 진행한다. 각 트위터 문장은 표 3의 전처리 대상에 대해 제거하는 과정을 거친다.

혐오 단어 사전은 국립국어원 비속어 사전과 게임 비속어 사전[7]과 청소년 언어의식 언어 실태 전국조사[8]의 부록인 비속어·은어·유행어 어휘 목록을 바탕으로 직접 구축하였다. 혐오 단어 사전의 수는 672개이다. 표 4는 혐오 단어 사전 정보이다. 혐오 단어 사전 정보 내 음절 n-gram의 수는 윈도우의 설정과 관련이 있다. 사전에 등록된 비속어의 최대 음절 길이가 6이므로 최대 6-gram까지 설정하였다.

표 3 전처리 대상
Table 3 The preprocessing object

The preprocessing object	Type
Twitter symbol	@(User name), RT(Retweet)
HashTag	Tag as starting with '#'
Symbol	A symbol excluded ./,/?/!
URL	Youtube or article URL and others
Emoticon	Emoji or special character
Phone number	The phone number that appears in promotional twitter sentence

표 4 혐오 단어 사전 정보
Table 4 Information about abusive word dictionary

n-gram	A number of word
2-gram	266
3-gram	263
4-gram	124
5-gram	14
6-gram	5

트위터 문장은 오타와 띄어쓰기 오류 등으로 포함하고 있는 비정형 데이터이기 때문에 정상적인 문장을 추출할 수 없으며 비정형 데이터는 자연어 처리를 작용하여 사용하기가 어렵다.

사용자들은 혐오 단어로 구성된 블랙리스트에 있는 단어 사이에 기호와 숫자 등을 사용하여 쉽게 블랙리스트를 매칭하는 방법으로 혐오 발언을 탐지할 수 없게끔 회피할 수 있다.

트위터 사용자가 작성한 하나의 트윗을 한 문장이라고 설정하였으며 한 문장을 음절 단위로 문장의 형태를 변경하여 초기 학습 코퍼스를 생성한다.

3.3 퍼지 범주 표현

퍼지 범주 표현은 명제 혹은 집합에서 참, 거짓과 같이 뜻이 명확한 것을 다루는 문제가 아닌 애매모호한 기준을 다루기 위한 것이다. 모호한 상태를 수식화하여 시스템을 구축하는 데 사용하고 기존의 0과 1로 이루어진 이진 논리의 한계를 극복할 수 있다.

우리가 생성한 실버 코퍼스(Silver corpus)는 혐오 단어 사전 매칭으로 부착한 레이블이며 이진 논리로 구성되어 있다. 하지만 혐오 단어 사전으로 부착한 레이블에 대해서 100% 신뢰하기 어렵다. 그렇기 때문에 트위터 문장이 어느 레이블에 속하는 정도를 표현하기 위해 퍼지 범주 표현을 적용한다.

퍼지 범주 표현을 적용하여 CNN 학습을 진행하고 CNN에서 softmax 함수의 값을 각 트위터 문장의 부착된 레이블에 대한 퍼지 범주 표현 값으로 설정하였다. 기존의 정답 레이블이 부착되어 있는 코퍼스를 이용하여 학습을 진행할 경우에 loss 계산은 정답 코퍼스의 레이블 값을 적용하여 계산한다.

퍼지 범주 표현 값을 적용하였을 때의 학습 방법은 이전 세대의 softmax 값을 각 문장에 대한 레이블 값이라고 정의한다. 이렇게 매 세대를 통해 이전 세대의

표 5 '블락 당한 거 자랑하는 양반도 대체 뭘 썼나 했더니 나 같으면 고소로 안 끝났을 듯'에 대한 퍼지 범주 표현
Table 5 Fuzzy category representation about 'The man who is proud of being blocked also what did write, if I was the same situation, I do not think it will end in a lawsuit.'

Generation	Reliability of Hate speech	Reliability of Non-hate speech
0	0.000000	1.000000
1	0.000068	0.999931
2	0.000173	0.999827
3	0.000382	0.999618

softmax 값을 정답 분류 값이라고 설정하고 loss 값을 통해서 학습이 진행되며 모델의 파라미터 또한 학습된다.

본 논문에서는 학습이 종료되고 학습 데이터가 재생성 되는 주기를 세대(Generation)이라고 정의한다. 한 세대는 학습을 진행할 때 Epoch과 상관없이 학습 코퍼스 문장 중 10%로 구성된 검증 코퍼스를 이용하여 성능 측정을 정확률로 설정하였으며, 최고 성능을 10번 갱신하지 못할 경우 한 세대의 학습이 종료된다.

트위터 문장에 대한 퍼지 표현 값을 생성하는 예시를 보여준다. 트위터 문장에 대한 퍼지화를 적용하기 위한 전처리된 예시 문장은 '블락 당한 거 자랑하는 양반도 대체 뭘 썼나 했더니 나 같으면 고소로 안 끝났을 듯'이다. '세대'의 값이 0이면 실버 코퍼스를 의미한다. 예시 문장이 혐오 발언 문장인지 비혐오 발언 문장인지를 분류하기 위한 학습을 진행한다.

여기서의 가정은 '문장 내 단어들이 혐오 단어 사전에 포함되어 있지 않다'이다. 그렇기 때문에 세대의 값이 0일 때 비혐오 문장에 대해 '1'라는 레이블이 부착된다. 퍼지 범주 표현을 사용함으로써 혐오 단어 사전에 매칭되지 않는 문장이라도 세대를 생성하며 학습을 진행할수록 '혐오 발언'에 대한 레이블 값이 생성된다. 표 5는 해당 예제 문장에 대한 퍼지 범주 표현 결과이다.

3.4 준지도 학습

준지도 학습 방법은 레이블이 부착되지 않은 대량의 데이터와 레이블이 부착된 소량의 데이터를 이용한다. 대부분의 준지도 학습 과정은 초기에 레이블이 부착된 데이터만으로 학습한 뒤 레이블이 부착되지 않은 데이터에 점차 레이블을 부착해나간다. 레이블이 부착된 데이터와 레이블이 부착되지 않은 데이터를 동시에 학습한다.

본 논문에서의 학습 코퍼스는 혐오 단어 사전으로 매칭하여 생성된 실버 코퍼스이기 때문에 준지도 학습을 적용하기 위해서 초기에 혐오 발언에 대한 레이블 9개의 트윗 문장과 비혐오 발언에 대한 레이블 11개의 트윗 문장을 사람이 직접 부착하였다.

퍼지 범주 표현을 적용할 경우 준지도 학습에 적용되지 않은 문장은 현재 학습 세대에서 이전 학습의 softmax 값을 사용하여 loss를 계산하지만 준지도 학습에 사용되는 20문장은 퍼지 범주 표현을 적용하지 않고 사람이 부착한 레이블 값을 이용하여 loss 값을 계산한다. 즉, 준지도 학습에 사용되는 20문장은 세대가 진행되어도 loss에 계산되는 레이블 값은 변화하지 않는다.

표 5의 예제 문장처럼 세대가 진행되면 레이블 부착 값은 변경되지만, 준지도 학습에 적용된 문장은 레이블 부착 값이 0 또는 1을 그대로 사용하도록 설정하였다. 그리고 주어진 데이터를 활용하여 평가 코퍼스에 대한 예측 모델을 구축하는 학습을 진행한다.

3.5 학습 모델

학습 모델은 CNN과 문장 벡터를 이용한다. CNN 모델은 두 가지 연산 층인 convolution layer, pooling layer와 최종적으로 fully-connected layer를 통해 분류를 수행하는 모델이다. 크기가 다른 다양한 필터를 사용하여 convolution layer에서 여러 개의 feature map을 생성한다. feature map의 값들을 activation 함수를 거친 후 max-over-time pooling 연산 단계를 거친다. max-over-time pooling은 각 필터의 여러 개 값 중에서 가장 큰 값을 선택하는 기법이다. 그 연산 값들이 fully-connected layer를 거쳐 출력이 결정되는 모델이다[9].

우리는 그림 2의 구조와 같이 CNN 구조에 추가적으로 문장의 정보를 얻고자 하였다. GRU는 Gated Recurrent Unit로써, RNN의 한 종류이다. LSTM의 장점을 유지하면서 계산 복잡도를 낮춘 셀 구조를 의미한다[10,11]. reset gate는 새로운 입력을 이전 메모리와

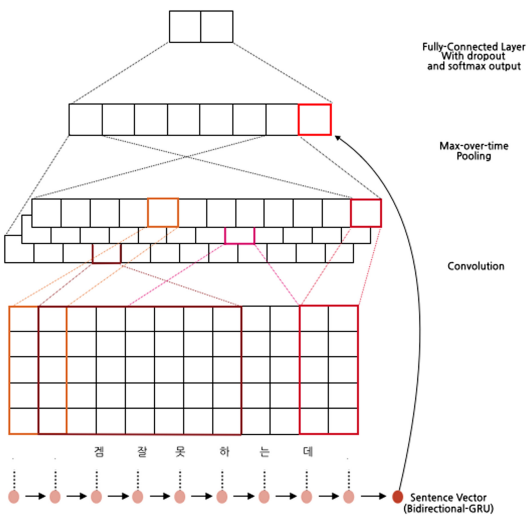

그림 2 학습 모델
Fig. 2 Learning Model

어떻게 조합하는지를 결정하며, update gate는 이전 메모리 정보를 어느 정도 유지하여 새로운 state를 계산할 것인지 결정한다. 따라서 CNN에 입력하였던 문장을 음절 단위로 Bidirectional-GRU에 입력하여 문장의 distributed representation에 해당하는 벡터를 구한다.

CNN 학습 내 convolution layer와 max-over-time pooling layer를 이용하여 그 문장에 대한 유용한 자질 벡터는 추출된다. 추출된 자질 벡터 뿐 아니라 추가적으로 문장 벡터를 사용하여 문장 정보를 얻고자 하였으며, CNN을 통한 자질 벡터와 문장 벡터를 concat하여 fully-connected layer를 통해 최종 혐오 발언 여부를 결정하는 softmax 분류기에 입력으로 하여 분류를 진행하고자 한다.

4. 실험 및 논의

실험은 크게 2개의 실험으로 분류되고 실험 1은 문장 벡터를 제외한 음절 CNN과 실험 2는 문장 벡터를 포함한 음절 CNN이다. 실험 2는 CNN의 네트워크 구조를 변경하여 세부 실험으로 구분된다.

2016년 3월 31일부터 2017년 8월 17일까지 작성된 트위터 문장들 중 벌통[12] 사용자가 검색한 총 13,882개의 키워드를 이용하여 트위터로부터 이 키워드에 매칭되는 트위터 문장들을 수집하였다.

본 논문에서 학습에 사용된 코퍼스는 혐오 단어 사전을 이용해 생성한 44,000문장이고 준지도 학습을 위한 사람이 레이블한 20문장이 포함되어 있다. 평가에 사용된 코퍼스는 수작업으로 생성한 13,082문장이다. 평가에 사용된 코퍼스 작성에는 총 5명이 참여하였고 중복 작업을 진행하지 않고 서로 다른 문장에 대해 2,600문장씩 레이블을 부착하여 생성하였다. 혐오 레이블 부착의 대상은 상대방을 비방하거나 욕설을 한다거나 성적인 주제로 이야기하는 경우 등을 '혐오 발언' 레이블을 부착하였다. 트위터 문장에 혐오 발언 여부에 대한 정답 코퍼스가 존재하지 않기 때문에 퍼지 범주 표현 값과 준지도 학습을 이용하여 학습 코퍼스를 생성하였다.

각 실험에 대한 정보는 표 6과 같다. Syllable CNN은 음절 단위의 1-layer의 convolution 단계를 거치는 모델이고 Syllable CNN2는 음절 단위의 CNN 내 구조가 2-layer의 convolution 단계를 거치는 모델이다. 문장 벡터는 bi-GRU을 이용하여 생성되고 max-over-time pooling 값과 concat하여 softmax 분류기의 입력으로 사용된다.

각 실험마다 파라미터를 다르게 사용하였으며 사용된 실험의 CNN 모델 파라미터 설정은 표 7과 같다. 필터 사이즈는 혐오 발언 사전에서 발생하는 혐오 단어의 최대 음절 길이까지 적용하였다.

표 6 각 실험에 대한 정보
Table 6 Information about each experiment

	Syllable CNN 1	Syllable CNN 2	Sentence vector
1	O		
2	O		O
3		O	
4		O	O
5	O	O	O

표 7 실험 1에 대한 파라미터 설정
Table 7 A setting of the parameters used in experiment 1

Parameters	Setting values
A count of filter	64
Filter sizes	2,3,4,5,6
Embedding dimension	50

표 8 실험 1의 세대별 성능
Table 8 Performance of experiment 1 for generation

Classify	Precision	Recall	F1 Score
Generation-1	65.40%	97.66%	78.34%
Generation-50	53.11%	97.94%	68.88%
Generation-100	52.01%	98.01%	67.96%
Generation-150	49.11%	98.01%	65.43%
Generation-200	29.76%	99.51%	45.82%
Generation-250	22.97%	99.86%	37.35%
Generation-300	21.91%	100.00%	35.94%

위의 설정으로 매 세대별 평가 코퍼스에 대해 측정하고 성능은 세대-300까지 설정하였고 50배수의 세대별의 성능을 측정하였다. F1 Score 성능은 다음 표 8과 같다.

세대가 생성될수록 음절 CNN의 모델은 혐오 발언 문장과 비혐오 발언 문장을 잘 구분하지 못하고 전체 평가 코퍼스 중 대다수의 문장을 혐오 발언 문장으로 예측하는 결과를 보였다. 이러한 문제점을 해결하기 위해서 본 논문에서는 음절 CNN뿐만 아니라 문장 벡터를 생성하여 이용한 모델로 실험을 진행한다. 사용된 실험 2의 CNN 모델 파라미터 설정은 표 9와 같다.

위의 설정으로 매 세대별 평가 코퍼스에 대해 측정하고 성능은 세대-300까지 설정하였고 50배수의 세대별의 성능을 측정하였다. F1 Score 성능은 다음 표 10과 같다.

표 9 실험 2에 대한 파라미터 설정
Table 9 A setting of the parameters used in experiment 2

Parameters	Setting values
A count of filter	64
Filter sizes	2,3,4,5,6
Embedding dimension	50
Sentence vector length	196

표 10 실험 2의 세대별 성능
Table 10 Performance of experiment 2 for generation

Classify	Precision	Recall	F1 Score
Generation-1	71.82%	97.42%	82.68%
Generation-50	73.99%	97.00%	83.95%
Generation-100	74.03%	97.20%	84.05%
Generation-150	75.07%	96.47%	84.43%
Generation-200	76.04%	96.47%	85.04%
Generation-250	77.20%	95.81%	85.50%
Generation-300	78.24%	95.04%	86.13%

표 11 사전 매칭과 실험 2의 세대-300을 기준으로 암시적 혐오 문장에 대한 성능 비교
Table 11 Comparison of performance for the implicit abusive sentences about matching dictionary and experiment 2 base on Generation-300

Classify	Precision	Recall	F1 Score
Matching dictionary	0.00%	0.00%	0.00%
Experiment 2	100.00%	6.09%	11.49%

본 논문에서 암시적 혐오 발언을 탐지하는 것이 목표이며 평가 코퍼스 내 암시적 혐오 문장은 82문장이다. 사전 단어 매칭으로 찾지 못하는 문장에 대해서 제안하는 방법으로 탐지가 가능하다. 사전 매칭과 실험 2 모델 중 세대-300의 성능 비교는 표 11과 같다.

혐오 단어가 포함되어 있지 않은 혐오 문장에 대한 성능 향상을 위해 실험 3은 음절 단위의 CNN 내 구조가 2-layer의 convolution 단계를 거치는 모델만 적용하여 실험한다.

실험 4는 실험 3과 동일한 CNN 모델과 문장 벡터를 연산하여 실험한다. 표 12와 표 13은 실험 3과 4의 세대별 평가 코퍼스에 대해 측정한 F1 Score 성능이며 표 14는 해당 실험 세대-300의 모델의 암시적 혐오 문장에 대한 성능 비교이다.

실험 3은 모든 문장에 대해서 “혐오 발언”이라고 레이블이 부착하는 결과로써 82문장에 대한 성능이 100%가 나왔으며 실험 4를 기준으로 분석하였다.

표 12 실험 3의 세대별 성능
Table 12 Performance of experiment 3 for generation

Classify	Precision	Recall	F1 Score
Generation-1	64.81%	97.70%	77.92%
Generation-50	69.71%	97.28%	81.22%
Generation-100	21.90%	100.00%	35.93%
Generation-150	21.90%	100.00%	35.93%
Generation-200	21.90%	100.00%	35.93%
Generation-250	21.90%	100.00%	35.93%
Generation-300	21.90%	100.00%	35.93%

표 13 실험 4의 세대별 성능
Table 13 Performance of experiment 4 for generation

Classify	Precision	Recall	F1 Score
Generation-1	71.60%	97.42%	82.54%
Generation-50	70.40%	97.35%	81.70%
Generation-100	73.47%	97.07%	83.64%
Generation-150	70.45%	97.10%	81.66%
Generation-200	68.36%	97.28%	80.30%
Generation-250	40.68%	97.63%	57.43%
Generation-300	27.39%	91.17%	42.42%

표 14 실험 3과 4의 세대-300을 기준으로 암시적 혐오 문장에 대한 성능 비교
Table 14 Comparison of performance for the implicit abusive sentence about experiments 3 and 4 base on Generation-300

Classify	Precision	Recall	F1 Score
Experiment 3	100.00%	100.00%	100.00%
Experiment 4	100.00%	59.76%	74.81%

2-layer로 구성된 CNN 네트워크는 전체 데이터에 대해 성능이 하락하지만 암시적 혐오 문장에 대한 성능은 상승하는 것을 알 수 있다. 이 결과로 분석해보자면 표면적인 문장 의미보다 내포된 문장 의미를 잘 예측한다고 할 수 있다. 표면적 의미를 잘 예측하는 1-layer CNN과 내포적 의미를 잘 예측하는 2-layer CNN을 함께 이용하는 모델을 생성하여 실험 5를 진행한다.

그림 3은 실험 5에 대한 모델 구조이다. 1-layer CNN

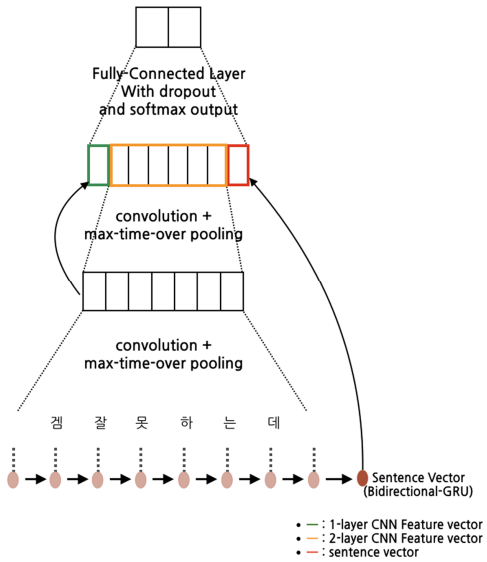

그림 3 실험 5의 모델 구조
Fig. 3 Model structure of experiment 5

표 15 실험 5의 세대별 성능
Table 15 Performance of experiment 5 for generation

Classify	Precision	Recall	F1 Score
Generation-1	69.85%	97.52%	81.40%
Generation-50	72.81%	97.28%	83.28%
Generation-100	74.61%	97.03%	84.36%
Generation-150	76.78%	96.72%	85.12%
Generation-200	66.78%	96.34%	78.88%
Generation-250	71.34%	95.99%	81.85%
Generation-300	77.62%	94.45%	85.21%

표 16 실험 5의 세대-300을 기준으로 암시적 혐오 문장에 대한 성능 비교
Table 16 Comparison of performance for the implicit abusive sentence about experiment 5 base on generation-300

Classify	Precision	Recall	F1 Score
Experiment 5	100.00%	14.63%	25.53%

을 이용한 자질 벡터와 2-layer CNN을 이용한 자질 벡터 그리고 문장 벡터를 concat하여 fully-connected layer를 통해 softmax 분류기의 입력으로 들어가서 분류 작업을 진행하는 모델 구조이다. 표 15는 세대별 모델 성능이며 표 16은 세대-300의 모델을 이용한 암시적 혐오 문장에 대한 성능 비교이다.

5. 결 론

자연어 처리 분야 중 혐오 발언 예측 시스템을 학습시키기 위한 코퍼스를 생성하는 것은 어렵다. 준지도 학습을 이용하여 온라인상에서 혐오 발언을 탐지하는 방법을 제안한다. 본 논문에서는 비정형 데이터인 트위터 문장에서 CNN과 문장 벡터를 이용하여 혐오 발언 여부를 분류하였다. 그리고 사전 매칭으로 생성된 실버 코퍼스 특성상 부착한 레이블에 대해 신뢰하기 어렵기 때문에 학습 코퍼스를 생성할 때 퍼지 범주 표현 값을 적용하였다.

이렇게 생성된 학습 코퍼스로 초기 분류기를 학습한다. 초기 분류기를 이용하여 트위터 문장에 대해 혐오 발언 여부에 대해 태깅한다. 이 때 softmax의 값을 퍼지 범주 표현이라고 하며 그 값으로 표현한다. 매 세대가 생성될 때마다 퍼지 범주 표현 값을 포함한 코퍼스가 생성되고 다음 단계의 학습에 이전 세대의 코퍼스가 입력이 되고 학습을 진행하고, 이 과정을 성능이 수렴할 때까지 반복한다.

기준 실험인 혐오 단어 사전 매칭보다 우리가 제안한 모델이 조금 높은 성능을 보였다. 실버 코퍼스는 단순 사전 매칭으로 정답을 생성한다. 따라서 문맥을 통한 혐오 발언 여부를 유추할 수 있는 정답을 반영하지 못한다는 문제점을 가지고 있다. 그러나 제안한 방법에서는 세대를 거치면서 혐오 단어 사전 매칭으로 탐지하지 못하는 문장을 혐오 발언으로 분류하는 것을 확인하였다.

표 17 세대-300을 기준으로 전체 실험에 대한 암시적 혐오 발언 문장에 대한 성능
Table 17 Comparison of performance for the implicit abusive sentence base on Generation-300

Classify	Precision	Recall	F1 Score
Matching dictionary	0.00%	0.00%	0.00%
Experiment 1	100.00%	100.00%	100.00%
Experiment 2	100.00%	6.09%	11.49%
Experiment 3	100.00%	100.00%	100.00%
Experiment 4	100.00%	59.76%	74.81%
Experiment 5	100.00%	14.63%	25.53%

혐오 단어 사전을 이용하여 실버 코퍼스를 생성하여 학습을 진행하였기 때문에 혐오 단어가 나타나면 '혐오 발언'이라고 결과를 내도록 학습이 될 것이라고 예상된다. 하지만 예상과는 다르게 준지도 학습과 문장 벡터로 인해 혐오 단어가 포함되어 있어도 '혐오 발언'이라고 결과를 내지 않는 것을 알 수 있다. 하지만 암시적 혐오 발언 데이터에 대해서는 세대를 생성할수록 혐오 탐지를 하지 못하는 결과를 나타낸다. 암시적 혐오 발언은 단어보다는 문장의 전체적 의미에 초점을 맞추어야만 한다. 게다가 음절 CNN과 문장 벡터의 조합만으로는 문장 전체적 의미에 초점을 맞추지 못했다는 점이 암시적 혐오 발언 탐지의 문제점으로 분석된다. 표 17은 세대-300을 기준으로 전체 실험에 대한 암시적 혐오 발언 문장에 대한 성능이다.

실험 1과 3은 모든 문장에 대해 혐오 발언 문장이라고 예측하였기 때문에 성능이 100%가 나왔으며 실험 4에 대해서 혐오 단어가 포함되어 있지 않은 혐오 문장에 대한 성능은 높지만 전체 평가 데이터의 성능이 42.12%이다.

모든 실험 중 실험 5의 경우가 전체 성능에 대해서는 실험 2의 성능이 조금 낮지만 혐오 단어가 포함되어 있지 않은 문장에 대해서는 가장 높은 성능을 보임을 알 수 있다. 이 결과는 다중 CNN이 문장의 표면적인 의미보다 내포적인 의미를 잘 찾는다라고 분석할 수 있다.

References

[1] [Online]. Available: https://ko.wikipedia.org/wiki/Cyber-bullying
[2] Bo-Bae, Lee, "Review the Legislative Direction to Regulate Hate Speech in Cyberspace," *KHU Global Business Law Review*, pp. 219-244, 2016.
[3] Björn. Gambäck, Utpal. Kumar. Sikdar, "Using Convolutional Neural Networks to Classify Hate-

Speech," *Proc. of the First Workshop on Abusive Language Online*, pp. 85-90, 2017.
[4] Nemanja. Djuric, Jing. Zhou, Robin. Morris, Mihajlo. Grbovic, Vladan. Radosavljevic and Narayan. Bhamidipati, "Hate Speech Detection with Comment Embeddings," *Proc. of the 24th International Conference on World Wide Web*, pp. 29-30, 2015.
[5] Ho-suk Lee, Hong-rae Lee, Yo-sub Han, "Semi-Supervised Learning Based Slang and Abusive Detection System," *Proc. of the KISS conference*, pp. 224-226, Vol. 2017, No. 6, 2017.
[6] Lei Gao, Alexis Kuppersmith, and Ruihong Huang, "Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach," pp. 774-782, 2017.
[7] Korea Game Industry Promotion Agency, "Study on the Guideline for Restoration of Game Language," 2008.
[8] Hanyang University Industry-Academy Collaboration Foundation, "National Language Survey of Language-Awareness of Youth," National Korean Language Publications, 2011.
[9] Yoon. Kim, "Convolutional Neural Networks for Sentence Classification," *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746-1751, 2014.
[10] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "Onthe properties of neural machine translation: Encoder-decoder approaches," *Proc. of SSST*, 2014.
[11] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Proc. of NIPS*, 2014.
[12] [Online]. Available: http:/eos.nlp.wo.tc, 2018-08-31.

박 다 솔

2014년 창원대학교 학사. 2017년 창원대학교 석사. 2017년~현재 창원대학교 친환경해양플랜트 FEED공학과(정보통신·컴퓨터전공) 박사. 관심분야는 자연어처리, 딥러닝, 기계학습

차 정 원

숭실대학교(학사). 포항공과대학교(석사, 박사). USC/ISI(박사후연수). 2004년~현재 창원대학교 컴퓨터공학과 교수. 관심분야는 자연어처리, 기계학습, 정보검색