준지도 학습 심층 신경망을 이용한 트위터 혐오 발언 문장 탐지

박다솔[0], 신창욱, 신영태, 차정원
창원대학교
dasol_p@changwon.ac.kr, papower1@changwon.ac.kr, zzz1421@naver,com, jcha@changwon.ac.kr

Detection of Abusive Sentence on Twitter Using Deep Neural Network with Semi-supervised Learning

Da-Sol Park, Chank-Uk Shin, Young-Tae Shin, Jeong-Won Cha
Changwon National University

요 약

사회관계망 서비스에서 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. 한국어 트위터 문장에 대해 혐오 발언 여부에 대한 레이블을 부착한 코퍼스를 생성하였다. 트위터 코퍼스 4만4천문장을 학습하였고, 1만3천여문장을 평가하여 정확률 80.50%, 재현율 95.25%, F1 Score 87.26% 성능을 얻었다. 논문에서 제안한 방법을 이용하여 사이버불링을 탐지하기 위한 방법으로 사용할 수 있다.

1. 서 론

청소년의 따돌림 문제가 심각한 사회문제로 대두되는 최근 스마트폰 보급과 인터넷 사용이 일상화되면서 학교 폭력은 사이버상의 따돌림과 괴롭힘으로까지 확대되고 있으며 사이버불링의 발생률이 증가하고 있다.

'사이버불링'은 사이버 범죄의 다른 말로, 인터넷과 관련된 기술상에서 다른 사람에게 피해를 입히는 모든 행위를 의미한다[1]. 사이버불링의 피해자는 자살을 선택할 정도로 심각한 물리적, 정신적 피해를 받는다. 그 중 언어폭력과 따돌림, 괴롭힘 등과 같은 혐오 발언으로 인한 사이버불링의 비중이 매우 높다. 블랙리스트를 이용하여 혐오 단어를 매칭할 수 있다. 하지만 단순히 혐오 단어 매칭만으로 탐지하기에는 한계가 있다. 혐오 단어 매칭만으로 탐지하지 어려운 문장의 예제는 아래와 같다.

(1) 싹 다 고소당하면 변호사 선임할 돈들은 있으세요?
머리에 든 거 보니 손에 쥔 것도 없어 보여서
(2) 기다란 침을 네 목에 꽂아 관통시켜볼까?

본 논문에서는 학습 데이터의 정답 코퍼스가 존재하지 않기 때문에 레이블이 부착되지 않은 대량의 데이터와 레이블이 부착되어 있는 소량의 데이터를 이용하는 준지도 학습(Semi-Supervised Learning)을 이용한다. 트위터 문장과 같은 비정형 텍스트에는 사용되는 단어가 더욱 다양한 유형으로 나타나기 때문에 분석을 위해 CNN(Convolutional Neural Network)과 문장 벡터를 혼합한 네트워크를 제안한다.

2. 관련 연구

[2]는 CNN을 이용하여 혐오 발언을 분류하는 방법을 제안했다. 단어 임베딩과 문자 n-grams를 사용하여 모든 단어에 대해 자질 임베딩을 생성한다. 단어 임베딩은 word2vec과 랜덤 벡터를 사용하는 두 가지 방법을 사용한다. 자질 임베딩은 단어 임베딩을 문자 n-gram vectors와 연결하여 생성된다. 분류는 인종 차별 발언과 성 차별 발언, 둘 다 포함한 발언, 혐오가 아닌 발언 총 4가지로 분류한다. word2vec을 사용한 시스템의 F1 score가 78.29%로 가장 높았다.

[3]은 신경 언어 모델을 사용한 혐오 발언 탐지를 제안했다. CBOW모델을 기반으로 한 paragraph2vec을 이용하여 문장과 단어를 저 차원 벡터로 임베딩한다. 임베딩된 결과를 이용하여 분류기를 학습시킨다. 추론을 할 때, 학습되지 않은 새로운 입력이 들어오면 이미 학습한 단어 임베딩을 삽입하여 추론한다. AUC(Area Under the Curve)로 측정한 성능은 0.8007이다.

[4]는 word2vec의 skip-gram을 사용한 비지도 학습과 블랙리스트, n-gram, edit-distance matric 그리고 외국어, 자소, 구두점 비윤리 탐지 기법을 사용한 지도학습을 결합하여 혐오 발언을 탐지하는 시스템을 제안했다. 비지도 학습으로 판별한 결과와 지도 학습으로 판별한 결과가 같을 경우 자동으로 비윤리 단어로 판별한다. 결과가

다른 경우 관리자가 직접 판별한다. 마지막으로 1종 오류를 최소화하기 위해 사전 filtering을 거친다. 성능은 F1 score 84.23% 이다.

3. 제안 방법

3.1 초기 학습 코퍼스 생성

트위터 문장은 오타와 띄어쓰기 오류 등을 포함하고 있는 비정형 데이터이다. 이러한 오류로 인해 비정형 데이터는 자연어처리를 바로 적용하기 어렵다. 따라서 우리는 띄어쓰기를 제거한 트위터 문장을 이용한다.

학습을 진행하기 위해서는 학습 코퍼스에 정답 레이블이 부착되어 있는 코퍼스가 필요하다. 그러나 정답 코퍼스 제작은 사람이 직접 생성해야 하기 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. 자체제작한 혐오 단어 사전의 수는 672개이다.

3.2 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 점차적으로 레이블을 부착해 나가며 레이블이 부착된 데이터와 레이블이 부착되지 않은 데이터를 동시에 학습한다. 본 논문에서는 초기에 혐오 발언 존재, 혐오 발언 비존재로 레이블이 부착된 각 10문장을 레이블이 부착되어 있는 데이터로 설정하여 학습을 진행한다. 본 논문에서는 CNN과 Bi-GRU와 카테고리 퍼지 표현(Fuzzy representation)을 이용하여 학습을 시도하였다.

퍼지 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. 우리는 혐오 단어 사전 매칭으로 silver corpus를 생성했기 때문에 부착된 레이블은 이진 논리이고 부착된 레이블에 대해서 100% 신뢰하기 어렵다. 그렇기 때문에 트위터 문장이 어느 레이블에 속하는 정도를 표현하기 위해 퍼지 표현을 적용하였다. 퍼지 표현의 값은 CNN 학습을 진행하고 CNN에서 softmax 함수의 입력값을 각 트위터 문장의 부착된 레이블에 대한 퍼지 표현 값으로 설정하였다.

이진 논리와 트위터 문장으로 형성된 학습 데이터는 학습을 통해 카테고리 퍼지 표현과 트위터 문장으로 변경된다. 본 논문의 세대(Generation)는 유전 알고리즘의 세대와 유사한 주기를 뜻하고 학습 데이터로부터 학습을 진행하여 퍼지 표현 값과 트위터 문장으로 학습 데이터가 재생성 되는 주기를 의미한다.

3.3 학습 모델

학습 모델은 CNN과 문장 벡터를 이용하고 그림 1과 같다.

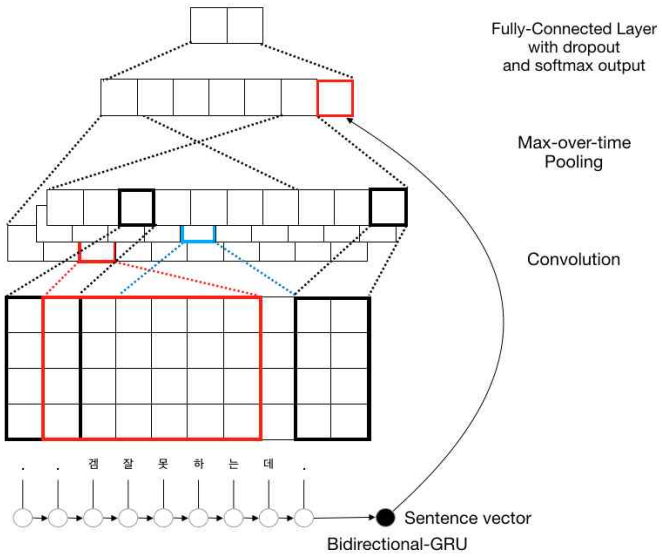

그림 1 학습 모델

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 [5].

우리는 위 CNN 구조에 추가로 문장의 정보를 얻고자 하였다. 따라서 CNN에 입력하였던 문장을 음절 단위로 Bidirectional-GRU에 입력하여 문장의 distributed representation에 해당하는 벡터를 구한다. 그렇게 구해진 문장 벡터는 max-over-time pooling이 수행된 후의 CNN 자질 벡터에 합쳐져 최종 혐오 발언 여부를 결정하는 softmax 분류기에 입력된다.

4. 실 험

본 논문에서 학습에 사용된 코퍼스는 혐오단어 사전을 이용해 생성한 44,000문장이고 평가에 사용된 코퍼스는 수작업으로 생성한 13,082문장이다. 트위터 문장에 혐오 발언 여부에 대한 정답 코퍼스가 존재하지 않기 때문에 퍼지 표현과 지도 학습을 이용하여 학습 코퍼스를 생성하였다. CNN 모델의 파라미터 설정은 표 2와 같다.

표 2 설정된 파라미터. 필터 사이즈는 사전에서 발생하는 혐오 단어의 길이이다.

파라미터	설정 값
필터 수	32
필터 사이즈	2,3,4,5,6
임베딩 차원	50

위 설정으로 기준 실험인 사전 매칭과 매 세대별 평가 코퍼스에 대해 측정한 F1 Score 성능은 표 3과 같다.

표 3 세대별 모델 성능

분류	Precision	Recall	F1 Score
사전 매칭	73.15%	**97.13%**	83.46%
세대-1	71.82%	97.41%	82.68%
세대-2	70.97%	97.38%	82.10%
세대-3	72.35%	97.38%	83.02%
세대-50	73.94%	96.93%	83.88%
세대-100	74.24%	97.07%	84.13%
세대-150	75.24%	96.30%	84.48%
세대-200	76.26%	96.40%	85.15%
세대-250	77.18%	95.29%	85.29%
세대-300	78.81%	95.01%	86.15%
세대-350	79.18%	95.81%	86.70%
세대-400	80.01%	95.85%	87.22%
세대-450	**80.50%**	**95.25%**	**87.26%**

세대별 모델에 대한 성능을 그래프로 나타내면 그림 4와 같다. 그래프에 X축은 세대명칭이고 Y축은 성능을 의미한다. 세대를 반복할수록 미세하지만 성능이 상승하는 경향을 보이고 있다.

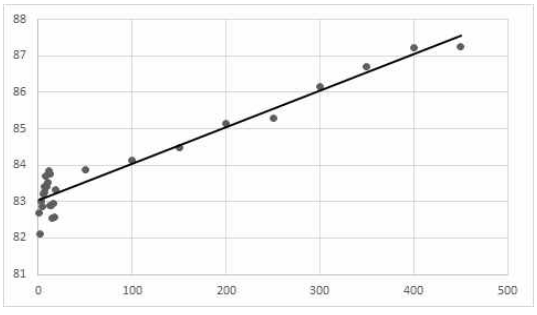

그림 4 세대별 성능 그래프

평가 코퍼스는 혐오 단어가 포함되지 않는 82문장을 포함하고 있다. 우리는 기계학습의 효과를 보이기 위해 이들에 대한 성능을 별도로 측정하였다. 표 5는 이들 문장에 대한 성능을 보여준다. 사전 매칭으로는 이 문장들을 찾을 수 없지만 제안 방법으로는 찾을 수 있음을 알 수 있다.

표 5 혐오 단어가 포함되어 있지 않은 혐오 문장에 대한 성능 비교

분류	Precision	Recall	F1 score
사전 매칭	0.0%	0.0%	0.0%
세대-400	100.0%	6.17%	11.63%

5. 결론 및 향후 연구

본 논문에서는 비정형 데이터인 트위터 문장에서 CNN과 문장 벡터를 이용하여 혐오 발언 여부를 분류하였다. 그리고 사전 매칭으로 생성된 silver corpus 특성상 부착한 레이블에 대해 신뢰하기 어렵기 때문에 학습 코퍼스를 생성할 때 퍼지 표현을 적용하였다.

기준 실험인 혐오 단어 사전 매칭보다 우리가 제안한 모델이 조금 높은 성능을 보였다. silver corpus는 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 사전 매칭으로 찾지 못하는 문장을 찾아내는 것을 확인하였다.

향후 연구로는 매 세대 모델 결과를 사람이 직접 레이블을 수정하는 active learning 기법을 적용해 볼 계획이다. 또한 일반적으로 CNN의 convolution layer가 깊어질수록 적절한 자질 추출을 한다. 따라서 convolution layer의 깊이를 추가하여 실험을 진행해 볼 계획이다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보 · 컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. 2017M3C4A7082524).

참고 문헌

[1] https://ko.wikipedia.org/wiki/사이버_폭력, 2017-10-25

[2] Björn. Gambäck, Utpal. Kumar. Sikdar. “Using Convolutional Neural Networks to Classify Hate-Speech”, Proceedings of the First Workshop on Abusive Language Online, pp 85-90. 2017.

[3] Nemanja. Djuric, Jing. Zhou, Robin. Morris, Mihajlo. Grbovic, Vladan. Radosavljevic and Narayan. Bhamidipati. “Hate Speech Detection with Comment Embeddings.”, In Proceedings of the 24th International Conference on World Wide Web, pp29-30, 2015.

[4] 이호석, 이홍래, 한요섭. “반자동 학습 기반의 비속어 및 욕설 탐지 시스템.”, 한국정보과학회 학술발표논문집, pp. 224-226. 2017.

[5] Yoon. Kim. “Convolutional Neural Networks for Sentence Classification”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746-1751, 2014.