워드 임베딩 기법을 이용한 혐오표현 데이터셋 자동 구축

서보성○ 김정민 강승식
국민대학교 소프트웨어학부
{sboseong,kimjm,sskang}@kookmin.ac.kr

Automatic Construction of Hate Speech Dataset by using Word Embedding Method

Seo Boseong○ Kim Jungmin Kang Seungshik
Dept. of Computer Science, Kookmin University

요 약

온라인상에서 자신의 의사를 자유롭게 표현할 수 있게 되면서, 특정 대상을 모욕하거나 차별 또는 언어 폭력을 행하는 등 혐오표현의 문제가 심각한 사회적인 문제를 유발하고 있다. 본 논문에서는 혐오표현으로 심각한 사회적 문제가 되고 있는 일베(일간 베스트) 커뮤니티 사이트로부터 수집된 댓글 데이터를 이용하여 워드 임베딩 모델을 학습하고 혐오성 키워드에 대한 워드 벡터로부터 혐오성 기준 벡터를 구성한다. 각 댓글에 대한 문장 벡터와 혐오성 기준 벡터와의 유사도 측정에 의해 특정 임계값 이상인 데이터들이 혐오표현이 뚜렷하게 나타나는 것을 확인하였고, 이를 이용하여 혐오표현 탐지를 위한 학습 데이터셋을 구축하였다.

1. 서 론

전 세계적으로 다양한 언어들에 대한 혐오표현을 방지하기 위하여 학습 데이터셋을 구축하고 있으며, 이를 기반으로 혐오표현 탐지 연구들이 수행되고 있다[1-4]. 본 논문에서는 특정 정치 성향, 지역, 성별 등의 혐오 현상이 강하게 나타나는 일베(일간 베스트 저장소)라는 커뮤니티의 댓글을 이용하여 혐오표현 탐지 데이터셋을 구축하였다. 혐오표현 데이터셋을 구축하기 위해 해당 커뮤니티의 댓글의 명사들로 학습한 Word2Vec 모델을 이용하였다. 단어 간의 연관성과 의미를 내포하는 Word2Vec을 이용하여, 혐오표현 Word2Vec 임베딩으로 부터 혐오표현과 유사한 Word2vec 임베딩을 가지는 댓글을 판별하여 혐오표현 데이터셋을 구축하였다. 이 데이터셋을 이용하여 게시물의 댓글이기 때문에 일반 평문과 일반적으로 접할 수 있는 혐오표현과 더불어 해당 커뮤니티에서만 사용하는 혐오표현을 탐지하였다.

2. 데이터셋 구축

2.1 데이터셋 구축 모델

워드 임베딩은 단어를 벡터로 표현하는 방법으로 각 단어들을 의미를 포함하는 벡터의 형태로 만들어준다. 해당 연구에서는 대표적인 워드 임베딩 중 하나인 Word2Vec을 이용하였으며, 학습에 필요한 데이터는 일베(일간 베스트 저장소)라는 커뮤니티의 댓글 중 명사만을 이용하여 학습시켰다. 명사의 경우 Konlpy의 Okt 형태소 분석기[1]를 이용하여 명사만을 추출하여, 총 1,196,345개의 댓글 중에 명사가 존재하지 않는 114,154개를 제외하고 1,082,191개 댓글의 명사들로 학습을 진행하였다. 각각의 댓글들은 평균 6.73개의 명사들로 구성되어 있으며, 가장 긴 댓글은 546개의 명사로 구성되어 있다.

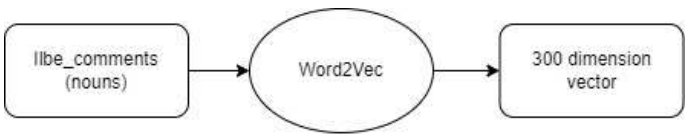

그림 1. 워드 임베딩

2.2 혐오 단어 사전 구성 및 혐오성 Score 계산

초기의 혐오성 시드(seed) 어휘를 '혐오'라는 단어로 시작하여 Word2Vec의 유사도 기반으로 확장을 진행하였다. 혐오성 어휘의 1차 확장 방법으로 다양한 유형의 혐오표현을 어휘집합에 포함시킬 수 있도록 임의로 5개 단어를 선택하였다. 혐오성 어휘집합은 1차 확장을 통해 얻은 어휘들을 기준으로 일정 유사도 이상의 단어들로 2차 확장을 진행하였다. 동일한 방법으로 어휘집합의 확장을 진행하였으며, 확장된 단어들을 혐오성 어휘집합으로 구성하였다. 구성된 혐오성 어휘집합의 각 단어들을 벡터화한 뒤 벡터합을 계산하여 혐오성 어휘집합에 대한 혐오성 기준 벡터를 계산하였다. 이를 도식화하면 그림 2와 같다.

1) https://konlpy.org/en/latest/

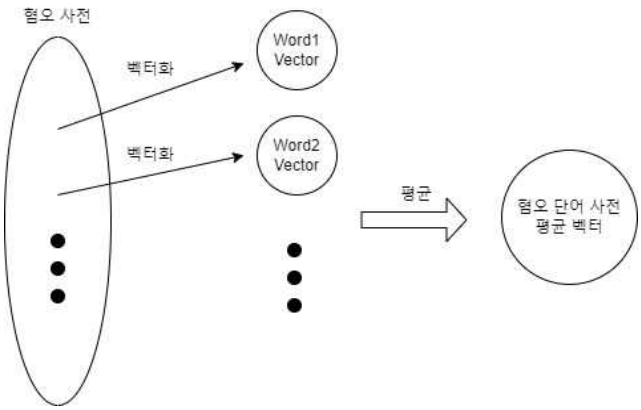

그림 2. 혐오성 어휘사전의 단어 벡터

마찬가지로 댓글 데이터의 각 댓글에 명사들에 대해서도 벡터화한 뒤 평균을 계산하여 각각 댓글 문장에 대한 문장 벡터를 계산하였다. 계산된 혐오성 기준 벡터와 댓글 문장에 대한 문장 벡터의 코사인 유사도를 계산하여 해당 댓글의 혐오성 유사도 값을 계산하였다.

3. 데이터셋 구축 및 실험

워드 임베딩 모델인 Word2Vec를 이용하여 초기의 혐오성 시드 어휘집합을 유사도 기반으로 확장하여 혐오성 어휘집합을 구성하였다. 이 어휘집합을 이용하여 각각의 댓글과 유사도 비교를 통해 혐오성 유사도 값을 계산하였다. 혐오표현 유사도가 계산된 댓글 중에 임의로 100개씩 10개 그룹을 랜덤하게 샘플링하여 확인한 결과 주로 0.35 이상의 유사도를 가지는 댓글에서 혐오표현이 발생하였다. 유사도가 그보다 낮아질수록 혐오표현의 발생이 줄어들고, 특히 0 미만의 유사도에서는 혐오표현이 포함된 댓글의 발생빈도가 현저하게 줄어드는 것을 확인하였다. 표 1은 혐오성 데이터로 판단되는 일베 데이터의 예이다.

표 1. 혐오표현 데이터 예제

하 시발 진짜 정신나간 새끼들 지금 극딜할게 널리고 깔렸는데 미친헛짓거리하고 자빠졌노
그러니까 왜 존재하면 안되는데
지랄 통베에서 따봉충까는게 노출처 통베똥먹는건데 이중성지리지?
한국인들은 그냥 병신임웰 하려고 하면 지랄지랄거리고이런 병신같은 민족사이에 박정희, 정주영, 이병철, 이건희,이명박이 탄생한게 기적임

표 2와 같이 혐오성 기준 벡터를 기반으로 혐오표현 학습 데이터셋을 구축하였다. 혐오성 기준 벡터와 일베 데이터의 유사도가 0인 데이터를 혐오표현이 아닌 'none' 라벨로 부착하여 비혐오성 데이터로 판정하고, 혐오성 유사도가 0.35 이상인 데이터를 혐오성 데이터로 판단하여 'hate' 라벨을 부착하였다.

표 2. 혐오성 데이터셋 구축 결과

유사도값	데이터 수	혐오성 라벨
0 미만	249,442	none
0.35 이상	136,617	hate

혐오성 데이터셋의 분류 정확도를 측정하기 위해 SVM(Support Vector Machine) 분류기 모델을 이용하였다. 자동 구축된 데이터셋에서 'none' 라벨 50,000개와 'hate' 라벨 50,000개를 임의로 추출하여 학습 데이터로 사용하였다. 정확도 측정을 위한 평가 데이터로는 Kaggle competition 중 'Korean Hate Speech Detection'[2]의 데이터를 사용하였다. 학습을 위한 파라미터의 경우 모두 default 값이며, 학습 데이터의 양은 100,000개, 평가 데이터는 5,679개로 진행하였다. 혐오성 분류 실험의 결과로 평가 데이터에 대해서 약 68%의 정확도를 보였다.

4. 결론 및 향후 연구

특정 정치 성향을 가지는 커뮤니티의 댓글들에 대하여 혐오성 어휘사전을 구축하는 연구를 수행하였다. SVM을 활용한 정확도 평가는 68%로 나타났는데, 이는 특정 커뮤니티에서만 사용하는 혐오표현들이 모델에 학습되었다는 제약과 명사 키워드로만 해당 문장을 평가했기 때문이다. 혐오표현 탐지에 대한 평가는 단순히 명사 키워드만을 혐오성의 속성으로 사용하였지만, 동사와 형용사 등으로 확장하고 띄어쓰기 오류 문제를 해결한다면 정확도가 높아질 것으로 예상된다.

참 고 문 헌

[1] 박승호, "혐오표현의 개념과 규제방법", 법학논총, vol.31(3), pp.45-88, 2019.

[2] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, Jun Long, "A Lexicon-based Approach for Hate Speech Detection, International Journal of Multimedia and Ubiquitous Engineering", Vol.10(4), pp.215-230, 2015.

[3] Jitendra Singh Malik, Guansong Pang, Anton van den Hengel, "A Comparative Study of Deep Learning Methods for Hate Speech and Offensive Language Detection in Textual Data", 2021 IEEE 18th India Council International Conference (INDICON), 2021.

[4] Sindhu Abro, Sarang Shaikh, Zafar Ali, Sajid Khan, Ghulam Mujtaba, Zahid Hussain Khand, "Automatic Hate Speech Detection using Machine Learning: A Comparative Study, (IJACSA) International Journal of Advanced Computer Science and Applications", Vol.11(8), pp.484-491, 2020.

2) https://www.kaggle.com/c/korean-hate-speech-detection