워드 임베딩 기법을 이용한 혐오성 어휘집합의 자동 확장

조단비, 강승식
신한은행 디지털혁신단, 국민대학교 인공지능학부
e-mail: daanv@shinhan.com, sskang@kookmin.ac.kr

Automatic Expansion of Hate Speech Expressions by using Word Embedding Technique

Danbi Cho, Seungshik Kang
Shinhan Bank, Kookmin University

요 약

국내외적으로 혐오표현의 심각성을 인식하고 있으며 이에 대처하기 위해 다양한 언어에서 혐오표현 탐지 연구가 진행되고 있으나 한국어 기반의 혐오표현 탐지 연구가 부족하다. 본 연구에서는 워드 임베딩과 딥러닝 기법을 활용하여 한국어 혐오표현 탐지 연구를 수행하였다. 한국어 문장의 맥락을 파악하고 단어의 의미론적 특성을 학습하기 위해 Word2vec과 FastText 기법을 활용하였으며 대용량 말뭉치를 사전 학습한 벡터를 구성하였다. 사전 학습한 벡터를 기반으로 두 차례의 어휘 확장을 수행하여 혐오표현 어휘집합을 구축하였다. 혐오표현 어휘집합은 각 단어에 대한 혐오 수준을 평가하며 문장 내 단어의 혐오 수준에 따라 문장의 혐오 수준을 판단하는데 활용된다.

1. 서론

디지털 서비스 확산을 통해 온라인 시장 규모가 확대되고 있으며 포털 사이트의 댓글과 소셜 네트워크 서비스(social network service, SNS)를 통한 의사소통이 활발하게 이루어지고 있다. 이러한 온라인 활동들은 익명 제도에 따라 이루어지고 있다. 익명 제도는 타인으로부터 사용자의 개인 정보를 보호해주는 역할로써 표현의 자유를 부여한다. 하지만 표현의 자유를 통해 불필요한 언어적 폭행이 이루어지고 있으며 이는 심각한 사회적 문제를 초래하였다.

편향성 문제를 해결하기 위해 어휘 편향성, 성 편향성, 정치적 편향성 등의 편향성 분류 연구가 수행되어 왔다[1-3]. 또한 혐오 표현 문제를 해결하기 위해 언어의 특성과 혐오 표현들의 특징을 분석하는 혐오 표현 탐지 모델이 연구되었다[4,5]. 혐오 어휘 사전 기반의 혐오 표현 탐지 방법론은 언어적 조합으로 생성된 새로운 혐오 어휘와 신조어를 추출하지 못하는 한계가 있다. 또한, 문장의 맥락을 파악하지 못한 상태에서 혐오 어휘 사전을 기반으로 하여 문장의 혐오 표현 여부를 판단하기 때문에 정확한 혐오 표현 탐지가 불가능하다. 이러한 한계를 극복하기 위해 딥러닝 기법을 활용한 연구들이 수행되었다[6,7].

영어, 아라비아어 등의 언어에 대하여 혐오 표현 탐지를 위한 연구들이 활발하게 진행되고 있다 [8,9] 하지만 한국어를 대상으로 한 혐오 표현의 데이터셋 구축과 모델 학습 연구가 부족하다. 특히 한국어에 대하여 '혐오 표현'의 정의가 명확히 제시된 바가 없으며 사람마다 혐오 수준을 판단하는 정도가 다르기 때문에 정확한 정의가 불분명하다. 본 연구에서는 워드 임베딩 기법을 활용하여 혐오성 어휘집합을 자동으로 구축하였다. 대용량 말뭉치로부터 문장 간 문맥 정보를 학습하고 특정 어휘에 따른 벡터 연산을 통해 유사한 어휘집합으로 확장하였다. 워드 임베딩 기법으로는 Word2Vec과 FastText를 사용하였다.

2. 워드 임베딩 기반의 어휘집합 확장

워드 임베딩은 데이터를 연속된 벡터 공간에 표현하는 방법이며 대표적으로 Word2vec과 FastText가 있다. 이들은 신경망 기반의 분산 표현 방식이기 때문에 연산량을 감소시키고 각 어휘들의 벡터를 저차원 공간에 표현할 수 있으며 유사도 계산이 가능하다는 장점이 있다[10,11]. 따라서 대용량 말뭉치를 워드 임베딩 기법에 적용하여 구성한 사전 학습 벡터 모델로 유사도를 평가하여 혐오 어휘집합을 자동으로 확장하여 구축하고자 한다.

임베딩 기법에 따라 단어 집합의 구성 방식과 주변 문맥 정보를 학습하는 방법론이 다르기 때문에 본 연구에서는 Word2vec과 FastText를 각각 학습하였다. 사전학습을 위한 대용량 말뭉치로는 KCC 뉴스기사 말뭉치를 활용하였다. Word2vec과 FastText 각각에 KCC150을 학습하였으며 KCC150보다 말뭉치 크기가 큰 KCC460을 Word2vec에 적용하여 학습 말뭉치 크기에 따른 성능의 차이를 비교하였다. KCC150과 KCC460은 뉴스 기사를 수집하고 정제하여 구축된 대용량 말뭉치이다. 표 1은 임베딩 기법에 적용한

두 가지 대용량 말뭉치 KCC150과 KCC460의 원시 데이터 문장수와 어절수 정보이다.

표 1. 사전 학습을 위한 대용량 말뭉치 정보

	문장 수	어절 수
KCC150	11,961,347	150,705,457
KCC460	29,316,426	467,649,207

KCC150 말뭉치를 Word2vec과 FastText 임베딩 기법으로 학습하여 Word2vec_KCC150 벡터 모델과 FastText_KCC150 벡터 모델을 확보하였으며 KCC460 말뭉치를 Word2vec 임베딩 기법으로 학습하여 Word2vec_KCC460 벡터 모델을 확보하였다.

3. 혐오성 어휘집합의 확장

워드 임베딩 기법을 활용하여 어휘를 확장하기 위해 단순히 '혐오'라는 어휘만으로 추가적인 어휘를 구성하기에는 의미가 제한적이고 확장 범위가 좁다. 따라서 '혐오' 단어와 유사한 어휘들을 시작 키워드로써 활용하였다. 세 가지 벡터 모델(FastText_KCC150, Word2vec_KCC150, Word2vcec_KCC460)에 대하여 '혐오' 어휘와 유사도 0.5 이상인 어휘들을 우선적으로 추출한 후, 세 가지 벡터 모델의 유사도 평균이 가장 높은 상위 5개의 어휘를 키워드로 지정하였다. 지정된 키워드는 '혐오', '증오', '경멸', '혐오감', '인종주의'이며 표 2는 각 사전 학습된 벡터 모델 별 각 키워드의 '혐오'와의 유사도와 각 평균 유사도를 나타낸다. 지정된 5개 키워드를 시작으로 1차 어휘 확장과 2차 어휘 확장 단계를 수행하였으며, '혐오'와 각 키워드와의 유사도를 혐오 수준으로 간주하였다.

표 2. '혐오'와 유사도 높은 어휘

벡터 모델	증오	경멸	혐오감	인종주의
FastText_KCC150	0.67	0.62	0.66	0.54
Word2vec_KCC150	0.75	0.74	0.69	0.60
Word2vec_KCC460	0.78	0.65	0.58	0.73
평균	0.73	0.67	0.64	0.62

3.1. 1차 어휘 확장

1차 어휘 확장에서는 키워드 (혐오, 증오, 경멸, 혐오감, 인종주의)를 기반으로 추출하였다. 각 키워드들은 세 가지 사전 학습된 벡터 모델에서 구축된 사전의 단어들과 유사도 연산이 이루어지며 유사도가 높은 어휘들을 1차 확장 어휘로 구성하였다. 이 때 키워드들은 '혐오' 와의 유사도 값이 다르다는 점을 감안하여 '혐오' 와의 유사도를 가중치로 부여하였다.

표 3은 사전 학습된 벡터 모델 별로 키워드에 따라 확장된 어휘 예시를 보여준다. 가중치가 반영된 유사도, 즉 혐오 수준을 함께 나타내며 혐오 수준을 기준으로 상위 5개의 어휘와 그 혐오 수준을 함께 제시하였다. Word2vec 모델이 FastText모델보다 추출된 유사 어휘의 수가 더 많았다.

표 3. Word2vec_KCC150을 이용한 1차 확장 예시

Word2vec_KCC150		
혐오	증오	
증오: 0.76 경멸: 0.74 혐오감: 0.70 폭력적: 0.66 천박: 0.65	경멸: 0.59 적개심: 0.59 증오심: 0.58 복수심: 0.58 미움: 0.57	
경멸	혐오감	인종주의
증오: 0.58 멸시: 0.58 모멸: 0.55 혐오: 0.55 냉소: 0.54	불쾌감: 0.53 모멸감: 0.52 공포감: 0.52 적대감: 0.52 동정심: 0.51	국수주의: 0.45 파시즘: 0.45 식민주의: 0.43 국가주의: 0.42 민족주의: 0.42

3.2. 2차 어휘 확장

1차 어휘 확장 단계에서는 키워드의 유사도 값을 가중치로 반영한 반면, 2차 어휘 확장 단계에서는 1차 확장 어휘들의 유사도 값을 가중치로 반영하여 2차 확장 어휘집합을 구성하였다. 사전 학습된 임베딩 벡터 모델별로 1차 확장 어휘와의 유사도를 계산하고 해당 유사도에 1차 확장 어휘의 혐오 수준에 가중치를 곱하였다.

2차 어휘 확장 단계에서는 1차 확장 어휘의 유사도 값을 가중치로써 반영하기 때문에 계산된 혐오 수준이 상당히 낮아지는 경향이 발생한다. 따라서 혐오 수준이 0.3 이상인 어휘들을 추출하여 2차 확장 어휘로 구성하였다. 표 4는 키워드별 2차 확장된 어휘 수이다. 각 키워드로부터 생성된 1차 확장 어휘들을 활용하여 2차 확장 어휘집합을 구성하였으며, 각 키워드에 속하는 어휘수를 합하여 함께 제시하였다.

표 4. 키워드별 2차 확장 어휘 수

	FastText_KCC150	Word2vec_KCC150	Word2vec_KCC460
혐오	907 (460)	30,827 (4,027)	8,584 (2,154)
증오	10 (6)	4,537 (853)	671 (327)
경멸	0 (0)	1,408 (261)	36 (29)
혐오감	3 (3)	117 (68)	5 (5)
인종주의	0 (0)	8,584 (2154)	1,469 (404)
총합	920	44,065	10,765

표 5는 2차 확장 어휘집합의 예시이다. 유사도를 기반으로 생성한 확장 어휘들은 모두 혐오 수준을 나타내기 때문에 키워드 별 구분된 어휘집합이 아닌 하나의 통합된 혐오 어휘집합으로 결합하였다. 확장된 어휘집합에는 중복된 어휘들이 존재한다. 예를 들어, '경멸'이라는 어휘는 '혐오'와

‘증오’ 키워드 모두에서 확장된 어휘로 추출되었다 특정 어휘가 중복되어 혐오 어휘로 판단되는 경우 혐오 수준이 그만큼 높아지는 것으로 간주되므로 중복 어휘의 유사도 합으로 혐오 수준을 계산하였다.

표 5. Word2vec_KCC460을 이용한 2차 확장 예시

Word2vec_KCC460		
혐오	증오	
고정관념: 0.45 혐오범죄: 0.43 괄시: 0.43 선입관: 0.43 Speech: 0.43	적대감: 0.45 불신감: 0.39 김치녀: 0.39 적대심: 0.38 복수심: 0.38	
경멸	혐오감	인종주의
냉대: 0.37 모멸: 0.35 핍박: 0.35 질시: 0.35 여성혐오: 0.34	공포감: 0.35 굴욕감: 0.33 모욕감: 0.32 성적수치심: 0.31 모멸감: 0.30	증오: 0.42 인종차별적: 0.41 인종차별주의자:0.40 인종차별: 0.38 인종혐오: 0.37

표 6은 임베딩 모델별로 결합된 혐오표현 어휘집합 크기와 혐오 수준이 가장 높게 나타난 어휘의 예이다.

표 6. 임베딩 모델별 혐오표현 어휘집합의 크기

사전학습 벡터 모델	어휘 개수	혐오성 어휘 예시
FastText_KCC150	460	증오: 3.75 두려움: 3.42 적대감: 3.06 열등감: 2.67 죄의식: 2.65
Word2vec_KCC150	4,039	경멸: 74.62 증오: 67.70 죄의식: 66.06 열등감: 63.30 증오심: 57.77
Word2vec_KCC460	2,164	인종주의: 33.03 혐오: 28.82 국수주의: 27.60 반유대주의: 24.04 극우주의: 19.76

4. 결론

워드 임베딩과 딥러닝 기법을 활용하여 한국어 혐오 표현 탐지 연구를 위한 목적으로 혐오표현 어휘집합을 자동으로 확장 구축하는 연구를 수행하였다. 문맥을 파악하고 단어의 의미 특성을 학습하기 위해 워드 임베딩 기법을 활용하여 대용량 원시 말뭉치로부터 사전 학습을 통해 임베딩 벡터를 구성하였다. 사전 학습한 벡터를 기반으로 두 차례의 어휘확장 기법을 수행하여 어휘집합을 확장하였다. 혐오성 어휘집합은 각 단어에 대한 혐오 수준을 평가하며 문장 내 단어의 혐오 수준에 따라 문장의 혐오 수준을 판단하는데 활용된다.

Acknowledgement

이 논문은 조단비 석사학위 논문[12]의 일부를 학술발표 논문으로 작성하였으며, 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2021R1F1A1061433).

참고문헌

[1] Cho, W. I., J. W. Kim, S. M. Kim, and N. S. Kim, “On measuring gender bias in translation of gender-neutral pronouns,” in Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing, pp.173-181, 2019.

[2] Bolukbasi, T., K. W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings,” in Proceedings of the 30th International Conferenc on Neural Information Processing Systems, pp.4356-4364, 2016.

[3] 조단비, 이현영, 정원섭, 강승식, “부분 단어 토큰화 기법을 이용한 뉴스 기사 정치적 편향성 자동 분류 및 어휘 분석,” 정보처리학회논문지: 소프트웨어 및 데이터 공학, Vol.10(1), pp.1-8, 2021.

[4] Gaydhani, A., V. Doma, S. Kendre, and L. Bhagwat, “Detecting hate speech and offensive language on twitter using machine learning: an n-gram and tfidf based approach,” in arXiv:1809.08651, 2018.

[5] Gitari, N., Z. Zuping, H. Damien, and J. Long “A lexicon-based approach for hate speech detection,”International Journal of Multimedia and Ubiquitous Engineering, Vol.10(4), pp.215-230 2015.

[6] Davidson, T., D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Proceedings of the 11th International AAAI Conference on Web and Social Media, pp.512-515, 2017.

[7] Alshalan, R., and H. A. Khalifa, “Hate speech detection in Saudi twittersphere: A deep learning approach,” in Proceedings of the 5th Arabic Natural Language Processing Workshop, pp.12-23, 2020.

[8] Rizwan, H., M. Haroon, and A. Karim, “Hate-speech and offensive language detection on Roman urdu,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp.2512-2522, 2020.

[9] Moon, J., W. I. Cho, and J. Lee, “BEEP! Korean corpus of online news comments for toxic speech detection,” in arXiv:2005.12503, 2020.

[10] Mikolov, T., I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in Neural

Information Processing Systems, Vol.26, pp.3111-3119, 2013.
[11] Bojanowski. P., E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the Association for Computational Linguistics, Vol.5, pp.135-146, 2017.
[12] 조단비, 워드 임베딩과 딥러닝 기법을 이용한 혐오표현 탐지, 국민대학교 석사학위 논문, 2021.