기호화된 혐오 상징어 검색을 통한 혐오 표현 탐지

김유민[1], 이환희[2]
중앙대학교 응용통계학과[1], 중앙대학교 AI학과[2]
e-mail : kimym7801@cau.ac.kr, hwanheelee@cau.ac.kr

Hate Speech Detection with Searching Coded Hate Symbols

Yumin Kim and Hwanhee Lee
Chung-Ang University

Abstract

As coded hate symbols intended for hate expressions are newly generated on the web, an automatic hate speech detection system that can reflect the meaning of new coded hate symbols is necessary. In this paper, we propose a novel pipeline for hate speech detection infused with knowledge about the meaning of coded hate symbols using Google API search. Moreover, we introduce the first hate speech detection dataset focusing on coded hate symbols with step-by-step explanations. Our pipeline provides higher performance with 68.14% accuracy compared to the detection that does not explain the meaning of coded hate symbols with an accuracy of 67.8%.

I. 서론

온라인 소셜 미디어에서 혐오 표현은 만연하게 사용된다. 이와 함께 혐오 표현 탐지의 중요성은 계속해서 강조되고 있다[1]. 혐오 표현은 읽는이로 하여금 정신적 악영향을 미칠 수 있기에, 사람이 이러한 혐오 표현을 직접 검열하는 것을 대신하고자 수많은 자동화된 혐오 표현 탐지 시스템 연구가 진행되어 왔다.

한편, 최근 거대 언어모델의 급격한 발전과 함께 자연어이해 및 생성 분야에서의 놀라운 성능이 여러 차례 입증되어 왔다. 그러나 이러한 거대 언어모델 조차도 혐오 표현 탐지에서는 저조한 성능을 보인다. 특히나, 온라인 혐오 표현과 같이 문맥을 참고할 수 없는 단일 문장이 큰 비중을 차지하는 경우 언어모델을 통해 혐오 표현을 탐지하는 것이 보다 어렵다. 또한, 혐오 표현 속 편향이 표면적으로 드러나지 않고 내재되어 있는 내재적인 혐오 표현의 경우, 현재까지의 언어모델은 그 내재된 의미를 명확히 파악하여 혐오 표현으로 탐지하는데 있어 낮은 정확도를 보이고 있다.

본 연구에서는 내재된 혐오 표현 탐지의 정확도를 높이고자, 검색 엔진을 통해 외부 지식을 활용하는 방법을 제안한다. 검색 엔진을 통해 문장 속에 내재된 혐오 상징어들의 의미를 파악하고, 언어모델이 이를 활용해 혐오 표현 탐지의 정확도 및 결과에 대한 해석 가능성을 높인다.

II. 본론

2.1 내재적인 혐오 표현 탐지

혐오 표현은 인종, 성별, 민족성, 종교, 국적, 성적 지향, 장애 또는 서열 등에 기초하여 혐오를 전달

또는 조장하는 표현이다. 내재적인 혐오 표현 (implicit hate speech)은 혐오 표현의 하위분류로, 반어법, 은유 등의 간접적인 언어 혹은 기호화된 언어를 사용하여 특정 대상을 비하하거나 대상에 대한 편향된 견해를 전달하는 발언을 의미한다[2].

내재적인 혐오 표현은 의미가 직접적으로 드러나지 않아서 자연어처리 시스템이 이를 정확히 탐지하는데 어려움이 있다. 특히 모델 학습에 필요한 내재적인 혐오 표현 데이터셋[3, 4, 5]은 그 절대적인 개수가 적고, 내재된 의미에 대한 명확한 설명[6]을 통하여 모델의 추론 및 탐지 능력을 강화하기 위한 설명 데이터셋 및 탐지 방법론에 관한 연구 또한 부족하다.

2.2 사용-인용 편향 완화

혐오 표현에 대한 대항 표현 (Counterspeech)에서 혐오 단어의 사용과 인용에 대한 구분을 통해 혐오 단어에 대한 무조건적인 민감도를 완화시킬 수 있다[8]. 이와 유사하게, 본 논문에서 제시하는 파이프라인은 기호화된 혐오 상징어가 포함된 문장일 경우 혐오 표현이 아님에도 불구하고 혐오 표현으로 오분류하는 경향성을 완화한다.

2.3 기호화된 혐오 상징어 검색을 통한 혐오 표현 탐지 파이프라인

Implicit Hate 데이터셋[4] 오분류 케이스 분석에 따르면, 기호화된 혐오 상징어 (Coded Hate Symbols)[7]가 사용된 경우와 상식이 필요한 경우가 총합 26%로 가장 큰 비중을 차지했다. 시간의 흐름에 따라서 새로운 기호화된 혐오 상징어는 계속해서 생성되므로, 언어모델에 새롭게 생성되는 혐오 상징어의 의미를 입력하고 업데이트하는 추가적인 과정이 필요하다. 이를 위해, 본 논문에서는 Google API를 활용하여 혐오 상징어의 의미를 검색하고, 검색된 의미를 혐오 표현 탐지 과정에서 언어모델에 설명하는 아래와 같은 파이프라인을 제시한다.

① 기호화된 혐오 상징어 검색

American Jewish Committee의 Translate Hate[9]와 위키백과를 주요 검색 도메인으로 특정하여, 기호화된 혐오 상징어의 의미를 Google API로 검색한 Top K=1 결과를 사용한다. 탐지 타겟 텍스트에 포함된 기호화된 혐오 상징어의 의미를 설명하기 위해 해당 검색 결과를 탐지에 사용할 언어모델의 프롬프트에 추가한다.

② 혐오 표현 탐지

2-shot learning을 활용하여, 기호화된 혐오 상징어가 포함된 혐오 표현 문장과 기호화된 혐오 상징어가 포함되었음에도 불구하고 혐오 표현이 아닌 문장 각각 1개씩을 예시로 프롬프트를 구성한다. 탐지 결과는 1) 혐오 대상 그룹, 2) 내재된 의미에 대한 설명, 3) 탐지 결과를 하나의 JSON 형태로 출력한다.

③ 기호화된 혐오 상징어의 사용-인용 구분

기호화된 혐오 상징어가 포함된 경우 혐오 표현이 아님에도 불구하고 혐오 표현으로 오분류하는 편향된 경향성을 고려하여, 혐오 상징어가 사용된 것인지 혹은 단순 인용된 것인지 구분한다. 혐오 표현 탐지 결과 혐오 표현으로 예측된 데이터만을 대상으로, 언어모델은 사용 혹은 인용에 대한 분류 결과를 출력한다. 사용으로 구분될 경우 기존의 탐지 결과인 "Hateful"을 유지하고, 반면에 인용으로 구분되는 경우 최종 탐지 결과를 "Not Hateful"로 변경한다.

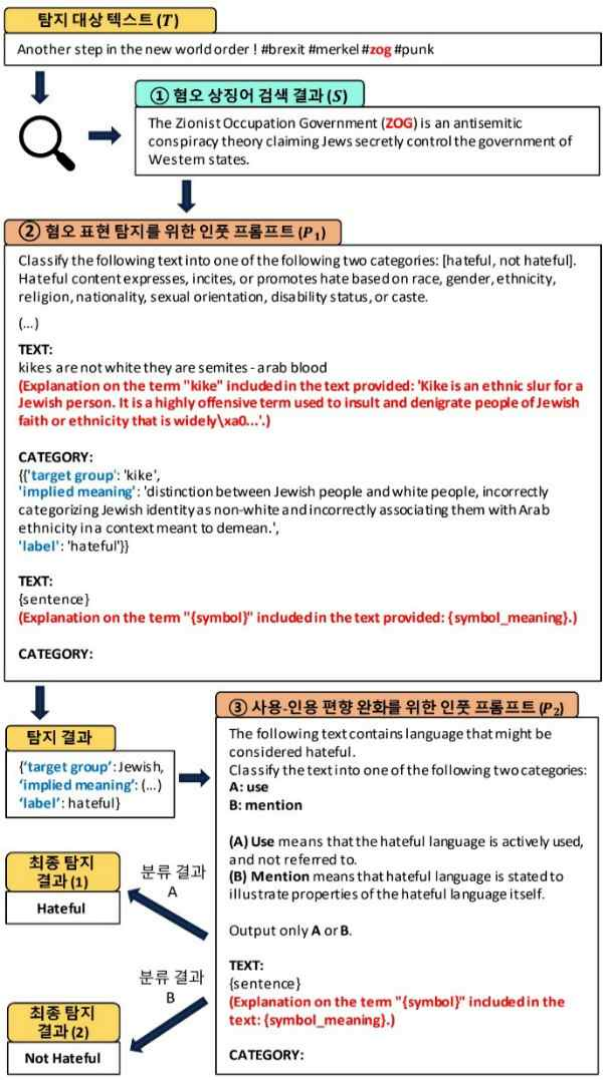

그림 1. 혐오 상징어 검색을 통한 혐오 표현 탐지 전체 파이프라인

Ⅲ. 실험

3.1 데이터셋

Implicit Hate 데이터셋[4]에서 기호화된 혐오 상징어가 포함된 295개의 데이터를 사용하여 테스트셋을 구축하였다. 기호화된 혐오 상징어 포함 여부는 American Jewish Committee의 Translate Hate 용어집[9]에서 정의한 혐오 상징어가 텍스트에 포함되어 있는지를 기준으로 분류하였다.

3.2 실험 세팅

본 연구에서는 실험을 위한 언어모델로 GPT-3.5-turbo-0125를 사용하였다. 우리는 GPT-3.5-turbo의 높은 수준의 추론능력을 기반으로[10] 혐오 표현에 대한 설명을 효과적으로 생성한다[11].

평가 지표로는 Accuracy (Acc), Precision (P), F1 score (F)와 Recall (R)을 사용한다. 혐오 표현으로 탐지된 경우인 "Hateful" 라벨을 1, 혐오 표현이 아닌 것으로 탐지된 경우인 "Not Hateful"을 0으로 표기한다. 예를 들어 P-0은 혐오 표현이 아닌 것으로 탐지된 라벨에 대한 Precision을 의미한다.

3.3 실험 결과

실험 결과, 기호화된 혐오 상징어를 검색하여 해당 단어의 의미에 대한 설명이 탐지 과정에 포함된 경우, Accuracy (68.14%)와 "Hateful" 라벨에 대한 F1 score (68.87%)가 가장 높았다. 이는 기호화된 혐오 상징어 검색을 통해 의미에 대해 언어모델에 설명하지 않은 경우와 비교하여 0.34%p 높은 결과이다.

Detection Setting	Acc	P-0	P-1	F-0	F-1	R-0	R-1
without Searching Hate Symbols							
2-shot	67.8	72.93	63.58	67.13	68.44	62.18	74.1
2-shot + Bias Mitigation	67.8	71.63	**64.29**	**68.01**	67.58	**64.74**	71.22
with Searching Hate Symbols							
2-shot	64.41	72.57	59.34	60.79	67.29	52.56	**77.7**
(Ours) 2-shot + Bias Mitigation	**68.14**	**73.48**	63.8	67.36	**68.87**	62.18	74.82

그림 2. 기호화된 혐오 상징어 의미에 대한 설명 포함 여부에 따른 혐오 표현 탐지 성능 비교

Ⅳ. 결론 및 향후 연구 방향

본 연구에서는 웹 상에서 변화하고, 새롭게 형성되는 기호화된 혐오 상징어에 관한 정보를 검색 엔진을 통해 언어모델에 정보를 제공함으로써, 혐오 표현 탐지의 성능을 높일 수 있는 파이프라인을 제시하였다. 본 연구에서 GPT-3.5-turbo-0125 모델의 2-shot learning을 수행한 결과, 기호화된 혐오 상징어의 의미를 언어모델에 입력한 후 혐오 표현을 탐지한 경우 68.14% Accuracy로 가장 높은 탐지 성능을 보였다.

본 연구는 사전 정의된 용어집 내의 기호화된 혐오 상징어를 포함한 데이터만을 대상으로 테스트셋을 구축하였다. 따라서 향후 연구에서는 전체 Implicit Hate 데이터셋[4]에서의 성능을 검증할 계획이다. 또한 Named Entity Recognition을 적용하여 검색 대상이 될 단어를 특정화하는 단계를 본 논문에서 제시하는 파이프라인의 가장 첫 순서에 추가할 예정이다. 이를 통해 사전 정의된 용어집 없이 텍스트 내의 모든 단어 중 검색 대상을 자동으로 특정하는 것이 가능해진다. 넓어진 검색 범위를 기반으로 향후 새롭게 생성되는 혐오 상징어에도 본 논문의 파이프라인이 효과적으로 작용할 수 있을 것으로 기대된다.

Acknowledgement

본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-01341, 인공지능대학원지원(중앙대학교)).

참고문헌

[1] Badjatiya, Pinkesh, et al. "Deep learning for hate speech detection in tweets." Proceedings of the 26th international conference on World Wide Web companion. 2017.

[2] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In Proceedings of the First Workshop on Abusive Language Online, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

[3] Kennedy, Brendan, et al. "The gab hate corpus: A collection of 27k posts annotated for hate speech." PsyArXiv. July 18 (2018).

[4] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 345–363, Online and Punta Cana, Dominican Republic.

Association for Computational Linguistics.

[5] Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

[6] Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023. Playing the Part of the Sharp Bully: Generating Adversarial Examples for Implicit Hate Speech Detection. In Findings of the Association for Computational Linguistics: ACL 2023, pages 2758–2772, Toronto, Canada. Association for Computational Linguistics.

[7] Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019. Learning to Decipher Hate Symbols. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3006–3015, Minneapolis, Minnesota. Association for Computational Linguistics.

[8] Gligoric, Kristina, et al. "NLP Systems That Can't Tell Use from Mention Censor Counterspeech, but Teaching the Distinction Helps." arXiv preprint arXiv:2404.01651 (2024).

[9] www.ajc.org/translatehateglossary

[10] Quyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.

[11] Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5490–5505, Singapore. Association for Computational Linguistics.