비윤리적 한국어 발언 검출을 위한 새 데이터 세트

박진원*, 나영윤**, 박규병**
*서울대학교 언론정보학과
**튜닙
jwp14812@snu.ac.kr, felix.ai@tunib.ai, ryan.ai@tunib.ai

A New Dataset for Korean Toxic Comment Detection

Jin Won Park*, Young-Yun Na**, Kyubyong Park**
*Dept. of Communication, Seoul National University
**TUNiB

요 약

최근 한국에서도 이루다의 윤리 이슈를 기점으로 딥러닝 모델의 윤리적 언어학습 필요성이 대두되었다. 그럼에도 불구하고 영어 데이터에 비해 한국어 데이터는 Korean Hate Speech Detection Dataset 이 유일하다. 이번 연구에서는 기존 데이터 세트의 유연성이 떨어지고 세부 라벨이 제한적이라는 문제를 개선한 새로운 데이터 세트를 제안하고, 해당 데이터 세트에 대하여 다양한 신경망 분류 모델을 적용한 벤치마크 결과를 공개한다.

1. 서론

올해 초 발생한 인공지능 챗봇 '이루다'의 성희롱 및 혐오 발언 논란으로 대화형 인공지능의 윤리적 언어 표현과 인식에 대한 문제가 대두되었다.[1] 인공지능은 자연언어로 된 사람들의 대화를 학습하는 과정에서 데이터에 내재되어 있는 각종 편향과 비윤리적 표현들을 그대로 흡수한다. 특히 익명성이 보장된 온라인 공간에서는 욕설, 혐오 표현 등 비윤리적 언어 사용이 보다 자유롭기 때문에 이러한 대화를 학습의 원재료로 사용할 경우 인공지능에도 필연적으로 편향이 생긴다. 따라서 무분별한 데이터 수용이 아닌, 문제 발언을 판별할 수 있는 윤리적 언어 학습의 필요성이 제기된다.

이에 본 연구는 대화형 인공지능의 비윤리적 발언 검출을 위한 한국어 학습 데이터 구축을 일차적 목적으로 한다. 욕설, 모욕, 폭력/위협, 외설, 범죄 조장, 혐오 표현 등 비윤리적 데이터를 걸러내는 세부 기준들을 새롭게 정의하고, 학습(train) 데이터 샘플 8 만 개, 검증(valid) 데이터 샘플 1 만 개, 테스트(test) 데이터 샘플 1 만 개를 구축하여 그 기준에 따라 분류하였다.

Bi-LSTM[2], GPT[3], ELECTRA[4] 등 여러 신경망 모델로 위 데이터 세트에 적용해 벤치마크로 삼았고, 성능 측정 지표로 MSE(Mean Squared Error)와 F1 score 를 사용하였다. 그 결과 모든 지표에서 ELECTRA 가 가장 높은 성능을 보였다.

2. 관련 연구

영어에 비해 기계 학습을 위한 한국어 비윤리 데이터 세트와 관련 연구는 아직 부족하다. 영어는 상대적으로 다양한 데이터가 공개되어 있으며 혐오 표현에 대한 기준이 세분화되어 있다. 통상적으로 개인이나 집단의 정체성을 표적 삼아 욕하거나, 공격하거나, 모욕하는 표현을 혐오 표현이라고 정의한다.[5] [6] 은 혐오 표현 데이터를 성별 혐오와 인종 혐오 등 세부 범주로 분류하고 [7]은 폄하, 반감, 위협 등 혐오를 표현하는 방식에 집중한다.

그러나 현재 공개된 혐오 표현 한국어 데이터는 국내 포털 사이트의 연예 기사 댓글을 수집한 Korean Hate Speech Detection Dataset[8]이 유일하다. 해당 세트에서는 댓글 94,000 문장을 Hate, Offensive, None 등 세 가지 기준으로 분류했으며 젠더 편향 포함 여부를 판별했다. 이 외에 인공지능의 윤리적 언어 학습에 대한 국내 연구는 비윤리적 언어 표현의 개념 정의와 검출 기준 설정에 초점을 두고 있다.[9][10] 따라서 인공지능이 실제 학습할 수 있는 데이터 구축이 반드시 필요한 실정이다.

하지만 대부분의 데이터 세트 연구는 비윤리적 표현의 유무를 이분법적으로 판별하기 때문에 검출 정확도나 데이터 활용의 유연성이 떨어질 수 있다. 나아가 특정 속성만을 강조한 분류 기준이 제한적일 수 있다. 본 연구는 보다 세분화된 데이터 분류 체계를 통해 기존 연구의 한계를 극복하고자 한다.

3. 데이터

3.1 데이터 수집

본 연구에서는 한국어 공개 데이터 세트 Korean Hate Speech Detection Dataset 의 200 만 개 unlabeled 데이터 중 16 만 개를 우선 사용하였다. 하지만 해당 데이터는 네이버 연예뉴스의 댓글로부터 수집했기 때문에 다양한 속성의 비윤리적 문장들을 확보하는 데 근본적 한계가 있었다. 우리는 이를 보완하기 위해 네이버 종교 뉴스[1] 2,000 건, 디시인사이드[2] 공익 갤러리 2,000 건, 일베사이트[3]에서 2,000 건, 성소수자 관련 네이버 뉴스[4] 2,000 건을 추가로 수집하여 사용하였다.

3.2 데이터 라벨링

두 명의 고용된 작업자가 각 문장에 대한 비윤리 속성과 정도를 판별했다. 각 작업자는 먼저 아래에 서술될 각 비윤리 속성별 정의와 측정 기준에 대한 가이드라인을 숙지하였다. 이후 본격적인 작업에 앞서 약 7000 개의 데이터를 직접 라벨링하며 연습하는 과정을 거쳤다. 이를 통해 일정 수준으로 합의된 결과가 나오도록 하였다.

5 단계 Likert 척도를 사용해 각 속성에 대해 0 부터 4 점까지 점수를 매겼다. 특정 속성이 문장에 없으면 0 점, 약하게 표현되어 있으면 1 점, 보통이면 2 점, 정도가 강하면 3 점, 매우 강하면 4 점을 주도록 했다. 점수 기준은 [11]에서 혐오의 강약을 구분한 방법을 차용했다. 비윤리적 표현이 간접적이고 공격성이 명확하지 않은 경우 또는 단순히 부정적 인식을 나타내는 경우에는 1~2 점을 줬다. 반면, 비윤리적 표현이 직접적이며 공격성이 명확하고 인격을 훼손하는 표현은 3 ~4 점을 줬다. 이에 대한 예시는 <표 1> 에서 확인할 수 있다

<표 1> 데이터 점수 별 예시(정치)

점수	예시
0	이야~ 정치계가 또뭘덮으려고 루머생성하냐
1	ㅈㄴ 무리수를 두면서 덮을려 하네 ㅋㅋㅋㅋㅋ
2	이게 다 대통령이 정치를 잘못한탓이다
3	좌빨성향 인간들 요즘 봄날이네. 천년만년 해쳐먹겠다
4	좌파 조국은 잘못절대 인정안하는데 우파는 쪽팔린거 알아서 잘못은 인정하지 좌파놈들 역겹다

3.3 데이터 라벨 속성

우리는 욕설, 모욕, 폭력/위협, 외설, 범죄 조장과 혐오 8 종 등 총 13 개로 비윤리 속성을 설정했다. 각 속성의 정의는 아래와 같으며 속성별 예시는 <표 2> 에서 확인할 수 있다.

욕설: 타인을 저주하고 인격을 훼손한 욕설 단어나 표현을 사용한 발화

모욕: 욕설을 사용하지 않지만 타인의 가치, 사회적 평가를 저하시키거나 경멸적인 감정을 표현하는 발화. 사실의 적시와 관계없이 타인을 폄하하는 경우 모욕으로 간주한다.

<표 2> 데이터 속성별 예시

속성	예시
욕설	구형 투싼 타고 다니는 쥐새뀌
모욕	남자 골퍼… 넘 못생겨서 몰입안됨
폭력/위협	여자끼리 싸우는걸 보니 보기 좋네요
외설	속 궁합도 나오는거야? 19 금?
범죄 조장	마약? 한 두번 해봐도 아무 악영향 없다.
성별	메갈들 방속에 좀 못나오게 해라
연령	틀닦들은 네이버 댓글 못갈게해라
인종/출신지	전북익산 출신인데 경주에서 사느라 지역감정 때문에도 힘들었을텐데 집안도 어려웠네
성적지향	동성애 소름끼쳐
장애	이런 저능아들은 도대체 뭔생각하면서 사는지 모르겠네
종교	개.독은 개.독을 알아보는 법!!`
정치성향	지금 정부는 박근혜때보다심한거같군덮을거있음
기타혐오	역시 개돼지들많네 막장이라고 욕할땐 언제고 잘도 처본다니들 수준이 딱 이수준이다

폭력/위협: 타인에게 물리적 또는 정서적 폭력을 가할 의도나 욕구를 표현한 발화. 나아가 타인에 대한 폭력을 지지하거나 폭력을 행사하도록 조장하는 발화.

외설: 성적 욕망을 자극하거나 성적 수치심을 주는 발화. 성적인 사실 관계를 묻는 것, 성적인 정보를 퍼뜨리는 것, 상대를 성적 대상화 하는 것, 성적 행위나 발언을 묘사하거나 강요하는 것 모두 포함한다.

범죄 조장: 범죄 행위를 조장하거나 지지하고 범죄 사건이나 범죄자에 대해 긍정적으로 평가하는 발화. 범죄 사건의 피해자에 대한 2 차 가해가 되는 발화도 포함한다.

혐오: 성별, 인종, 종교 등 정체성 요인을 이유로 개인이나 집단을 멸시, 모욕, 위협, 차별하거나 폭력을

[1] https://news.naver.com/main/list.naver?mode=LS2D&mid=sec&sid1=103&sid2=244
[2] https://gall.dcinside.com/board/lists/?id=gongik_new
[3] https://ilbe.com
[4] https://news.naver.com

선동하는 발화. 또한 혐오 사건, 행위, 발언을 지지하거나 부정하는 발화.

혐오 표현은 타인의 정체성 요인이 공격의 이유와 대상이 되는 점에서 다른 모욕적인 발화와 차이가 있다. 특히 혐오 표현은 편견과 차별을 양산하고 심각한 경우 증오 범죄로 이어질 수 있기에 다른 속성들보다 사회적 해악이 크다고 판단했다. 따라서 더 정확한 검출을 위해 8 개의 정체성 요인을 기반으로 세분화했다. **1) 성별, 2) 연령, 3) 인종/출신지, 4) 성적지향, 5) 장애, 6) 종교, 7) 정치성향, 8) 기타 혐오** (외모, 직업, 경제력 등 기타 정체성 요인에 대한 혐오).

3.5 교차 검증

데이터 세트 중 상대적으로 더 엄밀성이 요구되는 검증 및 테스트 데이터 세트들은 작업자들 간의 교차 검증을 실시하였다. 작업자들 간의 점수 차이가 3 점 이상 차이가 나는 결과물은 모두 제외시켰고, 2 점인 경우는 양 점수의 평균을, 1 점인 경우는 둘 중 더 큰 점수로 최종 라벨 결과물을 채택하였다.

3.6 라벨 분포

본 데이터의 클래스(욕설, 모욕, 종교 등)들은 상호 종속성을 지니는 경우가 많다. 예를 들어 욕설은 대부분이 모욕에 해당하는 식이다. 때문에 데이터 분포를 균등하게 만들기가 매우 힘들었다. 그리하여 이번 연구에서는 각 클래스당 데이터 수를 균등하게 맞추는 대신 각 클래스당 최소 600 개 이상을 수집하는 데 초점을 두었다. 각 속성(클래스)별 데이터 개수는 <표 3>과 같다.

<표 3> 속성별 데이터 개수와 비율(train, valid, test 를 모두 합한 수치)

속성	개수	비율(%)
모욕	57,806	41.5
기타 혐오	41,396	29.7
장애	14,900	10.7
욕설	7,263	5.2
성별	4,037	2.9
인종/출신지	3,316	2.5
외설	2,469	1.8
정치성향	2,297	1.6
연령	1,625	1.2
폭력/위협	1,617	1.2
범죄 조장	1,013	0.7
성적지향	838	0.6
종교	637	0.4
계	139,214	100

3.7 문제 부분 표시

문장에서 직접적으로 문제가 되는 단어나 구절을 { }로 별도 표시하여 더욱 구체적으로 비윤리적 표현을 검출할 수 있도록 했다.

예시: 후 나같은 {병신} 있냐

4. 벤치마크 실험

4.1 모델

여러 신경망 모델들을 사용해 본 데이터 세트의 품질을 검증하고 벤치마크 결과를 제공한다. 모델은 Bi-LSTM[2], GPT[3], ELECTRA[4]를 사용했다. Bi-LSTM 은 트랜스포머 구조가 나오기 전까지 문장 분류에서 좋은 성능을 거두었던 모델로 LSTM 계층에 역방향으로 처리하는 LSTM 계층을 추가한 모델이다. GPT 는 트랜스포머의 디코더 모델 구조를 사용한 모델로 다음에 올 단어를 예측하는 Autoregressive Language Modeling 방식을 사용해 자연어 생성에 매우 괄목할 만한 성능을 보였다. 마지막으로 ELECTRA 는 트랜스포머의 인코더 구조를 사용한 모델로서 학습 효율을 향상시키기 위해 Replaced Token Detection (RTD)이라는 새로운 사전 학습(pre-training) 태스크를 제안해 기존의 인코더 모델보다 더 좋은 성능을 기록했다. Bi-LSTM 의 경우 KcELECTRA[5]의 tokenizer 를 사용하였고, GPT 는 SKT 의 Kogpt2-base-v2[6]를, ELECTRA 는 공개된 체크포인트 중 TUNiB-Electra-ko-base[7]를 사용하였다.

<표 4> 모델별 파라미터 수 및 성능 결과

	파라미터 수	F1 score	MSE
Bi-LSTM	25.3M	0.953	0.04
GPT	125M	0.955	0.033
ELECTRA	110M	**0.963**	**0.029**

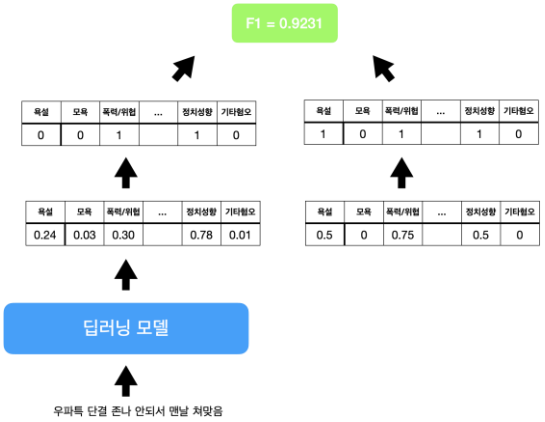
4.2 지표

데이터 세트의 특성상 Classification 과 Regression 을 같이 실험할 수 있다. Classification 은 F1 score 로, Regression 은 MSE(Mean Squared Error)으로 측정하였다.

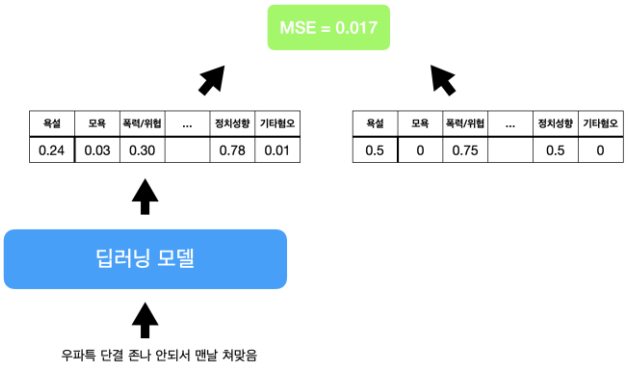
Metric 별 점수를 측정하는 방법은 (그림 1)과 (그림 2)에 보이는 것과 같다. 데이터 세트에 대한 학습을 수행하고 결과를 측정할 때 0 은 0, 1 은 0.25, 2 는 0.5, 3 은 0.75, 4 는 1 로 바꿔준 뒤 실험을 진행했다. Classification 의 경우 어떤 라벨의 속성이

[5] https://github.com/Beomi/KcELECTRA
[6] https://github.com/SKT-AI/KoGPT2

0.25 점 이상일 때 입력 텍스트의 속성이 해당 라벨에 속한다고 표현할 수 있다. 따라서 0 으로 라벨링된 데이터는 그대로 0 으로 놓고 0.25, 0.5, 0.75, 1 로 라벨링된 데이터는 1 로 놓았다. 모델이 예측을 0.25 이상으로 예측한 경우 1 로 인정하고, 0.25 미만으로 나타낸 경우 0 으로 내려 F1 score 를 측정하였다. <표 4> 을 보면 ELECTRA 모델이 GPT 보다 적은 파라미터를 가지고 있음에도 모든 지표에서 가장 좋은 성능을 낸 것을 볼 수 있다.



(그림 1) classification F1 score 계산



(그림 2) regression MSE 계산

5. 결론

이번 데이터 논문에서 우리는 Korean Hate Speech Detection Dataset 중 unlabeled 데이터와 디시인사이드, 일베저장소 등 온라인 커뮤니티와 포털 뉴스 댓글을 추가적으로 수집하여 총 10 만개의 데이터 세트를 구축했다. 또한 제한적이고 유연성이 부족한 기존 데이터 세트보다 더 광범위하며 조절 가능한 데이터 세트를 공개했다. 뿐만 아니라 이를 활용하여 다른 구조를 가진 여러가지 벤치마크들도 공개했다. 해당 연구가 한국어 혐오 문장 판별을 지원하고 윤리적 인공지능 개발을 위한 토대가 되기를 희망한다.

6. 사사(Acknowledgement)

이 논문은 과학기술정보통신부가 주최하고 과학기술정보통신부의 정보통신진흥기금으로 정보통신산업진흥원이 지원하는 개방형 경진대회 플랫폼 구축 사업의 '2021 년 인공지능 온라인 경진대회 우수 성과 기업 사업화' 사업지원을 받아 수행된 결과임 [과제번호: R-20210726-011600]

참고문헌

[1] 최세솔 & 홍아름,"AI 챗봇 '이루다'논란의 이슈 변화와 시사점", *전자통신동향분석*, 36(2), 93-101, 2021.

[2] M. Schuster & K.K. Paliwal, "Bidirectional recurrent neural networks", *IEEE Transactions on Signal Processing*, 45(11), 2673-2681, 1997.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. "Language models are unsupervised multitask learners", *OpenAI Blog*, 1(8), 2019.

[4] K. Clark, M. Luong, Q.V. Le, C.D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", *ICLR*, 2020.

[5] 박승호, "혐오표현의 개념과 규제방법", *법학논총*, 31(3), 45-88, 2019.

[6] Z. Waseem & D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on twitter", In *Proceedings of the NAACL Student Research Workshop*. ACL, San Diego, California, 2016, 88–93.

[7] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the worst: Dynamically generated datasets to improve online hate detection", In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. ACL, 2021, 1667-1682.

[8] J. Moon, W.I. Cho, J. Lee, "BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection". In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. ACL, 2020, 25–31.

[9] 이청호, 김봉제, 김형주, 변순용, 이찬규, "윤리적 인공지능을 위한 비도덕 문장 판별 온톨로지 구축에 대한 연구", *인공지능인문학연구*, 7(0), 149-170, 2021.

[10] 조태린, 김신각, 유희재, 김예지, 이주희. "대화형 인공지능의 윤리적 언어 표현을 위한 기 초 연구". 어문학, 140, 65-96. 2018.

[11] B. Vidgen & T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media", *Journal of Information Technology & Politics*, 17(1), 66-78, 2019.

[7] https://github.com/tunib-ai/tunib-electra