HateXplain
: A Benchmark Dataset for Explainable Hate Speech Detection

저자: Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, Animesh Mukherjee
발제자: 박채원
22-08-05

Abstract

- hate speech의 bias와 해석 가능성 측면에 대해 다룬 연구는 거의 없음
- hateXplain 제안
  - 세 관점에서 annotate됨 → hate, target(community), rationales(하이라이팅)
- 기존 SOTA 모델들이 설명 가능한 방법으로론 높은 점수를 내진 않음
- Rationales(근거)를 학습에 사용한 모델이 설명 가능성 측면에서 좋은 성능을 내며, 타겟을 향한 '의도치 않은 편향'에서 좋은 성능을 냄

Introduction

- 많은 모델이 일부 데이터셋에서 SOTA를 달성했다고 하지만, 일반화(generalize)엔 실패
- (혐오 의도 없이) 주로 공격 받는 정체성을 언급하는 것만으로 모델이 toxic으로 판단할 수 있음
  - -> 의도치 않게 편향된 예측
- 예측에 대한 설명이 부족
  - 혐오 발언 탐지 모델이 점점 더 복잡해지고 설명하기 어려워지고 있음
  - 이 연구에선 추가로 대상 분류와 근거를 함께 학습해 모델의 설명 가능성에 대해 접근

- 라벨링
  - 트위터와 갭에서 게시물을 수집하고, Amazon Mechanical Turk에 문의해 라벨링 진행
    - **증오, 모욕 또는 비방** 분류
    - 게시물에 언급된 **대상 커뮤니티** 선택
    - 분류 결정을 정당화할 수 있는 텍스트 부분 **하이라이팅** → “rationales” (근거)

Related work

- hate speech
  - 이전 연구들이 hate와 abusive/offensive를 혼동하는 경향이 있음
  - offensive이지만, hate와 동일시 될 수는 없는 메세지가 많이 있음
    - nigga ← 아프리카 아메리칸 커뮤니티에서 자주 사용됨
    - hoe, bitch ← 많은 랩 가사에서 사용됨
    - 이러한 단어들은 소셜 미디어에 널리 퍼져 있음

- Explainability/Interpretability
  - rationales(근거) 추출
    - Yessenalina, Choi, and Cardie → rationales 자동 생성 방법 구축
    - Lei, Barzilay, and Jaakkola → annotation 없이 질좋은 rationale 생성이 가능한 encoder-generator 프레임워크 제안
  - 해당 논문은 rationales(근거)의 개념을 사용하고, 인간 수준의 설명(작업자가 직접 하이라이팅 한 근거)을 포함하는 첫번째 혐오 발언 벤치마크 데이터셋을 제공

Dataset collection and annotation strategies

- Dataset sampling
  - 단일 lexicon(어휘집)을 사용해 게시글의 코퍼스 구축
    - 단일 lexicon → 3개의 연구에서 제시된 lexicon을 결합해서 하나의 lexicon 생성
  - 2019년 1월~2020년 6월의 기간동안 수집된 트윗에서 1%를 랜덤 샘플링
  - 링크, 사진, 비디오 등 추가 정보는 라벨링에 사용하지 않는다. *이모지는 제거하지 않음
  - 사용자 이름은 <USER> 토큰을 사용해 익명화

- Annotation procedure
  1. 텍스트가 hate인지, offensive인지, none인지 여부
  2. 대상 커뮤니티
  3. hate 나 offensive 판단 근거를 하이라이팅 하도록.

Dataset collection and annotation strategies

- 대상 커뮤니티 라벨링
  - 대상 커뮤니티 라벨링은 데이터셋을 풍요롭게 하기 위함이다.

Target groups	Categories
Race	African, Arabs, Asians, Caucasian, Hispanic
Religion	Buddhism, Christian, Hindu, Islam, Jewish
Gender	Men, Women
Sexual Orientation	Heterosexual, Gay
Miscellaneous	Indigenous, Refugee/Immigrant, None, Others

- 라벨링 설명서 제공
- 데이터셋 생성 단계
  - 파일럿 작업 (주석자 선정)
    - 20개의 게시물 ← hate 라벨링, 대상 라벨링
    - 621명의 작업자가 참여했고, 그중 253명이 선택됨
  - main 작업
    - 3명이 하나의 게시글에 라벨링을 진행하고 과반수 투표로 최종 선택
    - 결과물 → 트위터 9,055개 & 갭 11,093개
  - Krippendorff's alpha(작업자 간 동의) = 0.46 → 다른 혐오 발언 연구에 비해 높다.

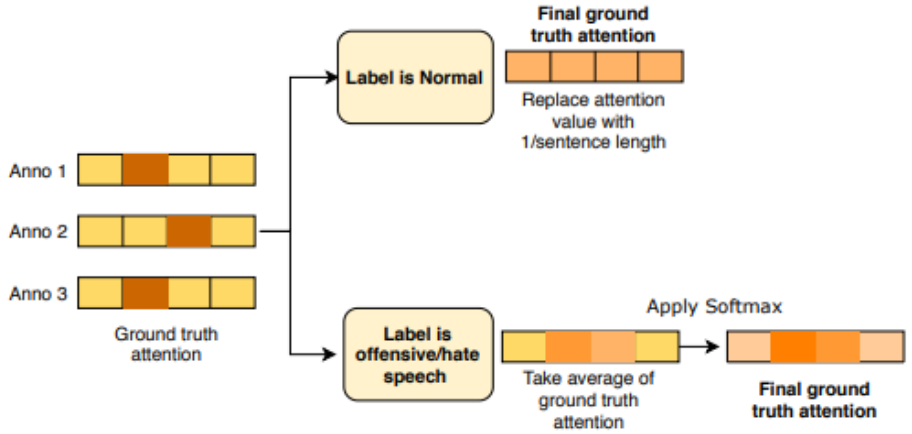
Dataset collection and annotation strategies

- 클래스 라벨
  - 3명의 작업자가 모두 다른 결론을 내린 919개의 게시글은 무시됨
  - +) 추가 연구로 이 데이터셋을 사용해 hate 대상에 관한 연구 가능
  - top 3 target
    - hate → 아프리카, 이슬람, 유대인
    - offensive → 여성, 아프리카, 동성애자
    - 이전 연구와 동일한 양상을 보임

  - 근거 하이라이팅
    - 2~3명의 작업자로부터 최종 라벨을 정당화하는 rationales가 제공되었다. (hate와 offensive만)
    - top 3 hate rationales → nigger, kike, moslems (30.02%)
    - top 3 offensive rationales → ratarded, bitch, white (47.36%)

Dataset collection and annotation strategies

- Ground Truth attention
  - 게시글을 attention vector로 변환 (rationales는 1로 표시됨)
  - 이는 문장의 토큰 개수와 같은 길이의 불리안 벡터이다.
  - 3개의 attention vector를 평균 취해 하나로 만든다
  - 기존 attention vector처럼 요소의 합을 1로 만들기 위해 softmax를 통해 정규화하여 생성
  - 문제점
    - Softmax 적용 시 rationales token과 non-rationales 토큰 사이 차이가 적을 수 있다.
    - 이를 해결하기 위해 **temperature 파라미터** 사용 → 확률 분포를 rationales에 집중시킴
  - none 게시물의 경우 각 요소간 동등한 분포의 Ground Truth attention를 가진다.



Metrics for evaluation

1. 성능 기반 방법
    1. 정확도, macro F1-score, AUROC로 hate 라벨의 분류기 성능 평가

2. 편향 기반 방법 (세가지 ROC-AUC)
    - “I love my niggas”가 흑인에 대한 혐오 발언이 아님에도 혐오로 분류됨
    - 이러한 모델의 의도하지 않은 정체성 기반 편향을 얼마나 줄일 수 있는 지에 대해 평가
    1. Subgroup AUC
        1. 테스트 셋에서 community를 ‘언급’하는 유해/무해한 게시글 선택
        2. 대상의 맥락에서 유해한 댓글과 정상적인 댓글을 분리하는 모델의 능력 측정
        3. 이 값이 높으면 모델이 특정 그룹에 대한 hate와 none을 잘 구별하고 있음을 보임
    2. Background Positive, Subgroup Negative(BPSN) AUC
        1. 테스트 셋에서 대상 언급 일반 게시글과 대상을 언급하지 않는 혐오 게시글 선택
        2. 대상에 대한 모델의 위양성 측정
        3. 대상 언급에 따른 성능 변화 관찰. 이에 따른 혼동이 있을 경우 의도하지 않은 편향이 있다고 판단

Metrics for evaluation

- Background Negative, Subgroup Positive(BNSP) AUC
  1. 테스트 셋에서 대상 언급 혐오 게시글과 대상을 언급하지 않는 일반 게시글 선택
  2. 대상에 대한 모델의 측정
  3. 대상 언급 혐오 게시글과 대상 언급 없는 일반 게시글을 혼동하지 않을 능력

- GMB (Generalized Mean of Bias) AUC
  - 편향 AUC의 멱평균 (멱평균 – 평균식을 일반화한 식)
  - 구글 대화 AI 팀이 캐글 대회에서 도입한 방법

$$M_p\left(m_s\right) = \left(\frac{1}{N}\sum_{s=1}^{N} m_s^p\right)^{\frac{1}{p}}$$

  - P → -5, N→서브그룹 개수, ms→ bias metric for subgroup, Mp→p번 거듭제곱 평균

Metrics for evaluation

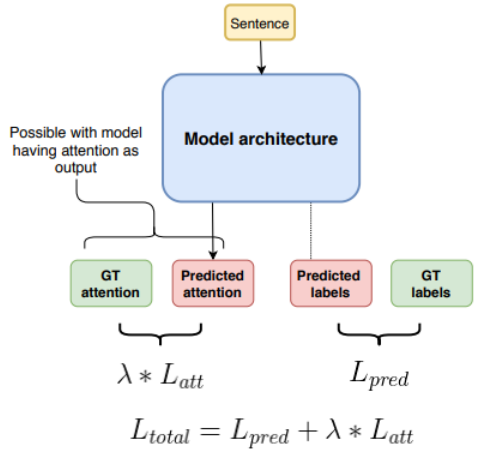
- 설명 가능성 기반 방법
- 타당성(plausibility)-해석이 인간에게 얼마나 설득력이 있는지 의미
  - discrete - IOU(Intersection-Over-Union) F1-score, token F1-score
    - IOU F1-score - 부분 일치에 대해 점수 할당. 토큰의 겹침 크기를 결합(union) 크기로 나눈 값
    - 토큰 F1-score - 토큰 수준의 정밀도와 재현율 측정 후 f1-score 도출
  - soft – AUPRC
    - 정밀도-재현율 곡선 아래 영역
- 충실성(faithfulness)-모델의 추론 과정에서 얼마나 정확하게 실제 reason을 반영하는지
  - 포괄성(comprehensiveness)
    - rationales(top 5)를 제거하면 모델 예측이 낮아질 것으로 예상
  - 충분성(Sufficiency)
    - 추출된 근거가 모델이 예측을 하기에 적절한 정도 측정
    - 작을수록 좋다

Model details

1.클래스 레이블만 사용해 학습

2.Attention, 클래스 레이블을 사용해 학습 (각각에 대해 attention을 나타내는 벡터를 출력해야함)

- CNN-GRU
  - 많은 혐오 발화 탐지 데이터셋에 SOTA를 달성한 모델
  - window size가 2,3,4(filter100 size)인 컨볼루션 1D 필터를 갖도록 수정
- BiRNN
  - sequential 모델(LSTM, GRU)에 임베딩 형태로 토큰을 전달
- BiRNN-Attention
  - 순차 레이어 후 attention layer를 가짐 → 컨텍스트 벡터 기반 어텐션 벡터 출력
  - attention layer 학습을 위해 예측된 attention layer의 출력과 ground truth attention을 CE 계산
- BERT
  - 최종적으로 ground truth처럼 단어에 attention을 줄 것이다.
  - BiRNN-Attention과 동일하게 교차 엔트로피를 사용하여 loss 계산

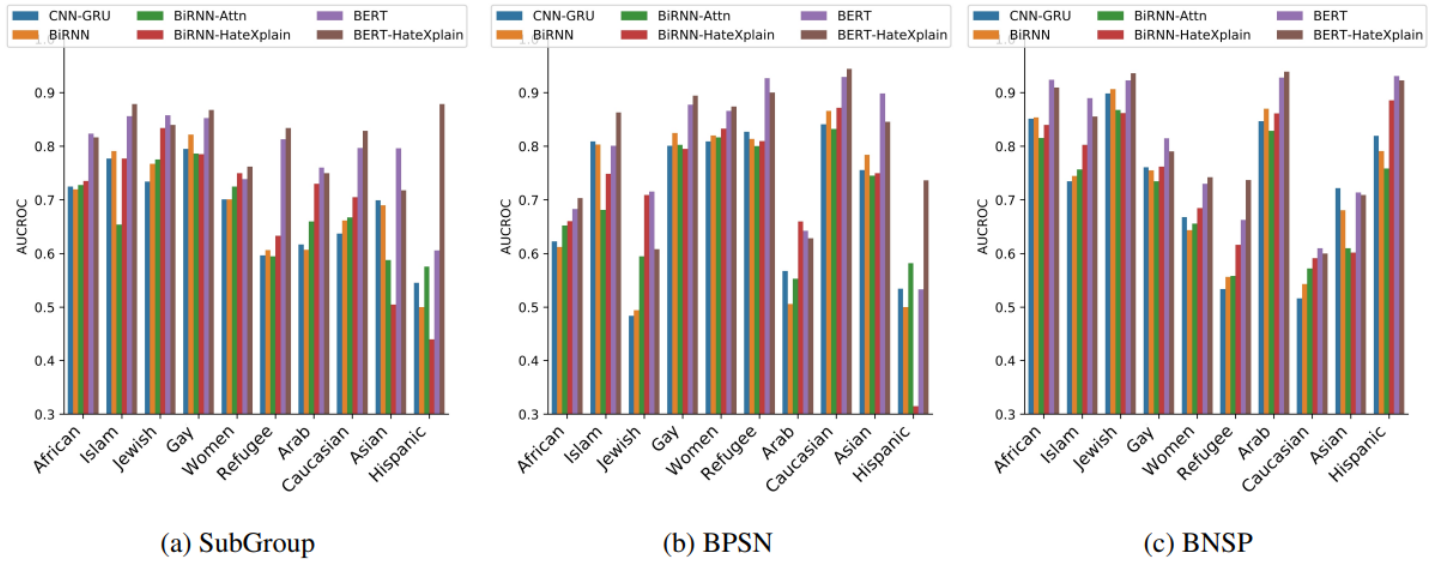


Result

Model [Token Method]	Performance			Bias			Explainability				
								Plausibility		Faithfulness	
	Acc.↑	Macro F1↑	AUROC↑	GMB-Sub.↑	GMB-BPSN↑	GMB-BNSP↑	IOU F1↑	Token F1↑	AUPRC↑	Comp.↑	Suff.↓
CNN-GRU [LIME]	0.627	0.606	0.793	0.654	0.623	0.659	0.167	0.385	0.648	0.316	**-0.082**
BiRNN [LIME]	0.595	0.575	0.767	0.640	0.604	0.671	0.162	0.361	0.605	0.421	-0.051
BiRNN-Attn [Attn]	0.621	0.614	0.795	0.653	0.662	0.668	0.167	0.369	0.643	0.278	0.001
BiRNN-Attn [LIME]	0.621	0.614	0.795	0.653	0.662	0.668	0.162	0.386	0.650	0.308	-0.075
BiRNN-**HateXplain** [Attn]	0.629	0.629	0.805	0.691	0.636	0.674	**0.222**	**0.506**	**0.841**	0.281	0.039
BiRNN-**HateXplain** [LIME]	0.629	0.629	0.805	0.691	0.636	0.674	0.174	0.407	0.685	0.343	-0.075
BERT [Attn]	0.690	0.674	0.843	0.762	0.709	0.757	0.130	0.497	0.778	0.447	0.057
BERT [LIME]	0.690	0.674	0.843	0.762	0.709	0.757	0.118	0.468	0.747	0.436	0.008
BERT-**HateXplain** [Attn]	**0.698**	**0.687**	**0.851**	**0.807**	**0.745**	**0.763**	0.120	0.411	0.626	0.424	0.160
BERT-**HateXplain** [LIME]	**0.698**	**0.687**	**0.851**	**0.807**	**0.745**	**0.763**	0.112	0.452	0.722	**0.500**	0.004

- performance
  - rationales를 사용한 모델이 성능 지표 측면에서 나은 성능을 보임 (- HateXplain 모델들)
    - explainability 성능 측정을 위해 중요 토큰을 사용했고, 이 rationales를 attention으로 뽑냐, LIME으로 뽑냐 가 token method
- Explainability
  - BERT-HateXplain[Attn]은 충분성에 대해 최악의 점수를 가짐
  - LIME이 attention에 비해 더 faithful한 결과를 내는 것 같다
- performance보다 타당성 및 충실도 점수가 선호되는 경우가 있을 것이기 때문에 HateXplain이 유용한 도구가 될 수 있음
- 람다(attention loss 조정 파라미터)의 값을 증가시키면 모델 성능, 타당성은 향상되나, 포괄성은 저하됨

Result



(a) SubGroup (b) BPSN (c) BNSP

- bias
  - performance와 동일하게 rationales를 사용한 모델이 의도치 않은 편향을 줄여준다는 측면에서 더 잘 수행된다.
  - AUROC 값이 커뮤니티마다 큰 편차가 있다.
  - BERT-HateXplain은 편향을 더 잘 처리한다.

Limitation and Conclusion and future work

- Limitation
  1. 추가적인 맥락 부족
     1. 프로필 정보, 성별, 과거 게시글 등 → 분류에 도움을 줄 것으로 예상
  2. 영어에만 국한된 연구

- Conclusion and future work
  - HateXplain → 혐오 발언 탐지를 위한 새로운 벤치마크 데이터셋. 20k개의 게시글로 구성
  - 3개의 annotation (혐오, 대상, rationales)
  - 여러 SOTA모델을 테스트 해보니 분류에서 좋은 성능을 낸 모델이 항상 그럴듯하고 믿음직한 근거를 제공하진 않는다.
  - 향후 연구로, 기존 혐오 데이터셋에 HateXplain을 적용하고자함

KOLD

- 카이스트 연구팀이 제안한 설명 가능한 한국어 모델을 만들기 위한 혐오 데이터셋
- 댓글과 함께 컨텐츠 제목을 제공해 맥락을 고려한 라벨링 진행
- 세 단계의 annotation 프로세스
  - 해당 댓글이 offensive 한가?
  - 대상 타입 분류 / 대상 세부 카테고리 분류
  - Offensive와 target의 Rationales 하이라이팅
- hateXplain 데이터와 비교했을 때, 가장 많이 나타나는 대상 top10이 완전히 다른 양상을 보임
  - 이를 통해 각 나라 문화에 따른 데이터셋 필요성 보임
- 세가지 실험 진행
  - 문장 분류 (hate, target, specific target)
  - Span 예측 (토큰 분류-NER, BIO)
  - 문맥 정보 제거 (제목 정보 제공 유무 비교)

현재 진행중인 연구

- 설명 가능한 혐오 탐지 모델을 만들기 위한 데이터셋 구축 중
- annotation 프로세스
  - 혐오 스케일링: Hate 항목(성, 직업, 장애, 지역/인종 등)에 대해 0-3 scale
  - 대상 분류: Hate 항목, 개인, 그 외 대상 에 대해 대상 분류
  - rationales 하이라이팅: 혐오와 대상의 rationales 하이라이팅
- Question answering 모델을 사용해 혐오를 탐지하는 방법 고려
- 작업자(annotator)의 정보를 통해 데이터 분석 가능성

- 추후 최종 데이터 확인 후 여러가지 분석 & 실험 진행 예정