Why Is It Hate Speech? Masked Rationale Prediction for Explainable Hate Speech Detection

Jiyun Kim, Byounghan Lee, Kyung-Ah Sohn

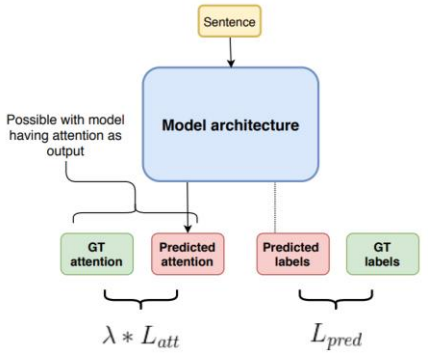
발표자: 박채원

22-11-04

HateXplain

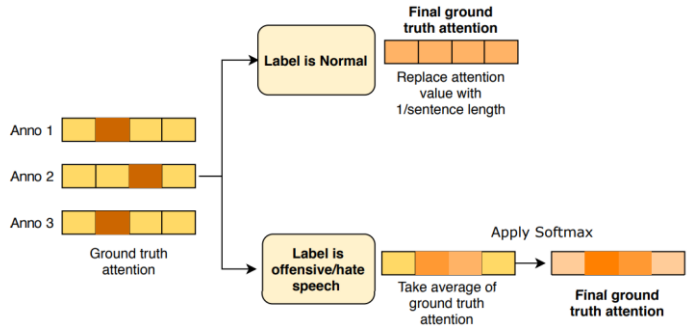
- 혐오 탐지 문제에 라벨(hate, offensive, none) 뿐만 아니라 rationale(혐오 판단의 근거 span)를 사용해 학습하는 방법을 제안
- HateXplain 데이터셋
  - 트위터, 갭의 게시물 20,148개
  - hate, offensive, none 로 3진 분류
  - 혐오 분류 중 라벨이 hate나 offensive일 경우 작업자가 혐오로 판단한 근거 span 하이라이팅
  - 3명의 작업자가 하나의 데이터에 라벨링 진행 (다수결 방법 사용해 최종 라벨 결정)
- rationale을 ground truth로 삼고 이를 모델의 attention weight과 cross entropy 취해 loss 계산

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

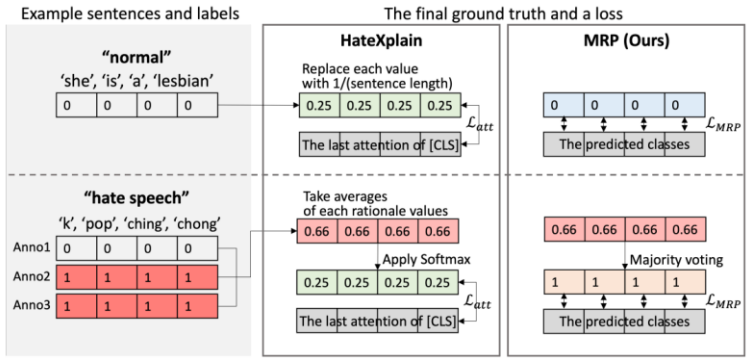


Introduction

- 차별과 편견이 빠르게 퍼져나가면서 그들의 인권을 해칠 수 있고 이로 인해 현실 범죄를 불러일으키기도 함. 그렇기 때문에 온라인에서의 혐오를 규제하는 게 중요해짐
- 분류 성능(정확도, f1 score 등..)에 더해 편향과 설명 가능성을 평가
  - 특정 단어가 포함되어 있을 때 모델이 편향된 탐지를 만드는 경향이 있음. 이는 차별을 강화할 수 있음
  - 그렇기 때문에 단어는 **맥락**에서 판단 되어야 한다. (혐오 탐지 문제에선 특히 중요)
- 해당 연구는 HateXplain데이터가 편향과 설명 가능성을 모두 고려한 혐오 데이터셋 이므로 이를 사용
- 기존 HateXplain에선 세 작업자의 rationale을 합하는 과정에서 모두 더하고 softmax를 취하는 방법을 사용했는데 이 경우 모든 토큰이 근거일 경우 normal과 같은 rationale을 갖게 되므로 다른 방법 취함



HateXplain의 방법



해당 논문에서 제시된 방법

Method

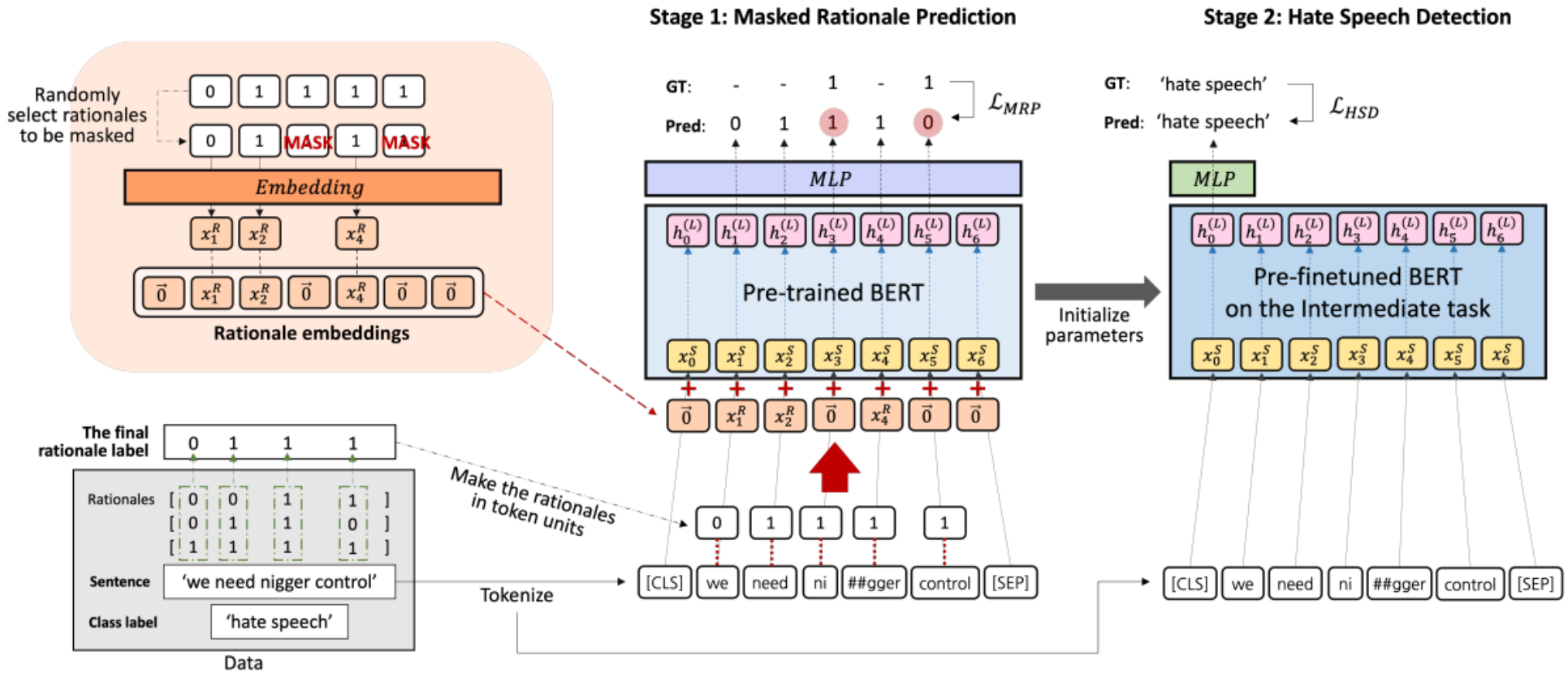
- MRP (Masked Rationale Prediction)
  - 주변의 가려지지 않은 rationale을 기반으로 가려진 토큰의 human rationale label을 예측
  - pre-training과 finetuning의 사이 task로 제안 (Pre-finetuning on an intermediate task)
  - 즉, BERT pretraining -> MRP -> Hate Speech Detection finetuning
  - 토큰 분류 문제에 기초함 (1: rationale, 0)
  - MRP는 MLM과 다름 -> 전체 토큰을 대상으로 마스킹 토큰을 선택하는 게 아니라 rationale 대상

- MRP 과정에서 학습된 파라미터가 혐오발언 탐지 task의 초기 파라미터가 됨
  - 인간의 근거 능력을 학습함으로써 모델의 편향, 설명 가능성 측면에서 성능이 향상될 수 있음.

- BERT-RP와 BERT-MRP 모델을 새롭게 제안하며 실험 진행
  - BERT-MRP: n%의 rationale만을 마스킹 후 이를 예측하도록 학습
  - BERT-RP: 전체 rationale을 마스킹 후 이를 예측하도록 학습 (n=100)
  - 두 모델이 좋은 성능을 냄(BERT-MRP가 SOTA를 달성)

$$H_{MRP}^{(0)} = X^S + \tilde{X}^R,$$
$$H_{MRP}^{(l+1)} = \text{Transformer}(H_{MRP}^{(l)}),$$
$$\hat{X}^R = \text{MLP}(H_{MRP}^{(L)}).$$

Method



- 혐오 발언 탐지(본 task) 학습에선 rationale이 사용되지 않고 단어 토큰만을 사용함

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
  - 각 편향 score(AUC)에서 그룹 멱평균(GBM)을 구해 제시

Metrics for evaluation

3. 설명 가능성 기반 방법
    - 타당성(plausibility)-해석이 인간에게 얼마나 설득력이 있는지 의미
        - discrete(0.5이상은 1, 미만은 0) - 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)
            - rationales를 제거하면 모델 예측이 낮아질 것으로 예상
        - 충분성(Sufficiency)
            - 추출된 근거가 모델이 예측을 하기에 적절한 정도 측정
            - 전체 문장을 갖고 예측한 확률에서 rationale만을 갖고 예측한 확률을 빼서 구함
            - 이 값은 작을수록 좋다

Models

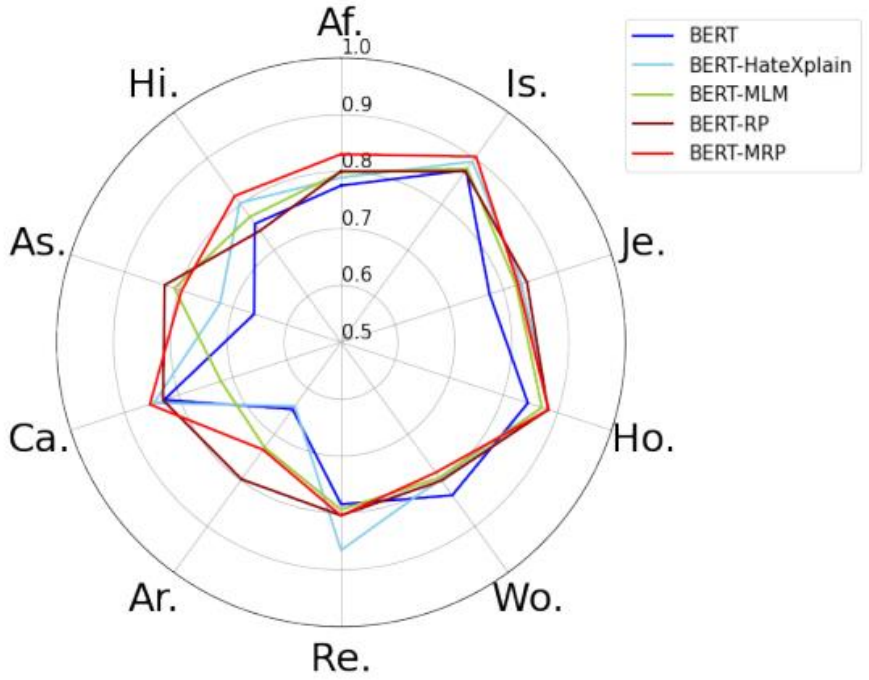
- BERT-base-uncased 모델 사용
- BERT-HateXplain
  - 최종 attention값(cls 토큰)을 GT로 사용하며, loss는 어텐션 로스와 분류 로스의 합
  - HateXplain 논문에서 제시된 것과 완전 동일
- BERT-MRP
  - pre-finetuning task의 효과를 보여주기 위해 평가됨
  - 50%의 라벨을 마스킹
- BERT-RP
  - 100%의 rationale 라벨을 마스킹
  - token classification
  - 그냥 rationale label을 전부 예측하는 것
- BERT-MLM
  - 문장에서 15%의 토큰을 마스킹

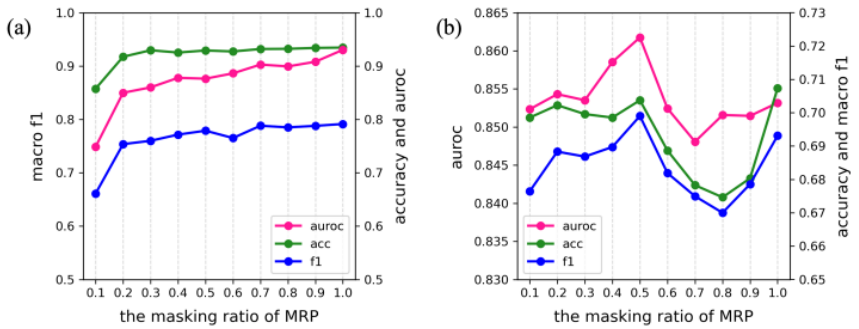
Experiments

Model			Performance			Bias		
	ration.	pre-fin.	Acc.	Macro F1	AUROC	GMB-Sub.	GMB-BPSN	GMB-BNSP
BERT			69.0	67.4	84.3	76.2	70.9	75.7
BERT-HateXplain	✓		69.8	68.7	85.1	80.7	74.5	76.3
BERT-MLM		✓	70.0	67.5	85.4	79.0	67.7	80.9
BERT-RP	✓	✓	**70.7**	69.3	85.3	81.4	74.6	84.8
BERT-MRP	✓	✓	70.4	**69.9**	**86.2**	**81.5**	**74.8**	**85.4**

Table 1: Results for the performance-based and the bias-based metrics. Scores in bold type are the best for each corresponding metric, while the underlined are the second best, and so are in Table 2.

Model			Explainability				
			Plausibility			Faithfulness	
	ration.	pre-fin.	IOU F1	Token F1	AUPRC	Comp.	Suff. ↓
BERT [Att]			13.0	49.7	77.8	44.7	5.7
BERT [LIME]			11.8	46.8	74.7	43.6	0.8
BERT-HateXplain [Att]	✓		12.0	41.1	62.6	42.4	16.0
BERT-HateXplain [LIME]	✓		11.2	45.2	72.2	**50.0**	0.4
BERT-MLM [Att]		✓	13.5	43.5	60.8	40.1	11.9
BERT-MLM [LIME]		✓	11.3	47.2	76.5	43.4	**-5.5**
BERT-RP [Att]	✓	✓	13.8	50.3	73.8	45.4	7.2
BERT-RP [LIME]	✓	✓	11.4	49.3	77.7	48.6	-2.6
BERT-MRP [Att]	✓	✓	**14.1**	**50.4**	74.5	47.9	6.7
BERT-MRP [LIME]	✓	✓	12.9	50.1	**79.2**	48.3	-1.2





Experiments

No.	Model	Example sentence and Rationale	Label
Case 1 Explicit hate speech			
1	Human	imagine unironically believing the high iq kike meme	HS
	BERT-MRP	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
	BERT-RP	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
	BERT-HX	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
2	Human	<user> ffs these ghetto blacks are ignorant like cracker mountain hillbillies	HS
	BERT-MRP	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
	BERT-RP	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
	BERT-HX	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
Case 2 Implicit hate speech			
3	Human	yah christian island republic shall not have queers or witches on it if we find any death	HS
	BERT-MRP	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	HS
	BERT-RP	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	HS
	BERT-HX	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	NO
4	Human	you can not culturally enrich a moslem	HS
	BERT-MRP	you can not culturally en ##rich a mo ##sle ##m	HS
	BERT-RP	you can not culturally en ##rich a mo ##sle ##m	NO
	BERT-HX	you can not culturally en ##rich a mo ##sle ##m	NO
5	Human	<user> he is infected with jihadi virus he will spread it to others	HS
	BERT-MRP	<user> he is infected with jihad ##i virus he will spread it to others	HS
	BERT-RP	<user> he is infected with jihad ##i virus he will spread it to others	HS
	BERT-HX	<user> he is infected with jihad ##i virus he will spread it to others	OF

Conclusion

- 편향과 설명가능성을 고려하는 혐오 발언 탐지 방법을 제안
- 사전학습된 언어 모델을 두 단계에 걸쳐 파인튜닝하는 프레임워크를 도입
- intermediate task로 근거 마스킹 예측, MRP를 제시
- 결론적으로 해당 모델이 탐지, 편향, 설명 가능성 측면에서 SOTA 성능을 달성
- 예측된 rationale이 모델의 굳건함을 증명함

- base 모델과 비교를 위해 BERT 모델만을 사용했지만 이 방법은 다른 트랜스포머 인코더 기반 모델에도 적용될 수 있다.