APEACH: Attacking Pejorative Expressions with Analysis on Crowd-Generated Hate Speech Evaluation Datasets

Kichang Yang, Wonjun Jang, Won Ik Cho

ENMLP 2022

발제자: 박채원

23-02-10

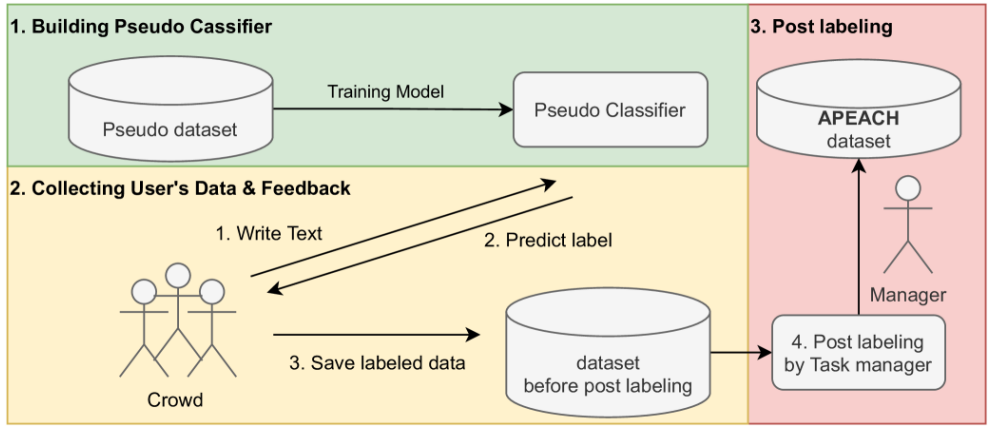
Abstract

- 혐오 표현 탐지에서는 다양한 도메인을 포함하는 학습, 평가 데이터를 구축하는 것이 중요
- 일반적으로 데이터셋 구축 scheme은 웹(소셜 미디어)에서 텍스트를 크롤링 하고 고용된 작업자가 이에 라벨링을 진행
- 하지만 이러한 관습적인 구축 방식에 의해 데이터가 싱글 도메인으로 한정되어 도메인 일반화가 부족한 경우가 있음 (ex 연예 섹션 뉴스 댓글만으로 구성되어 다른 도메인의 일반화에 어려움을 겪음)
- 도메인이 한정된 경우 도메인 중복으로 인해 성능이 과대평가 되기도 함

- 이러한 문제를 해결하고자 'APEACH' 제안
  - 익명의 작업자가 최소한의 post-labeling을 따라 혐오 발언 데이터를 만들도록 하는 것
  - 이 평가 데이터 구축 방식을 통해 사전학습 데이터와 평가 데이터 사이 어휘 중복에 덜 민감하게 모델의 성능을 적절히 평가할 수 있는 데이터를 구축할 수 있다.

Introduction

- 기존 혐오 표현 데이터셋 구축은 일반적으로 텍스트에 라벨을 다는 것으로 진행됨
- 하지만 이러한 방법은 데이터셋의 신뢰를 방해하는 몇가지 제한 사항이 있음
  - 온라인 자료의 특성에서 오는 라이선스 및 개인 정보 문제의 잠재적 위험성
  - 제한된 도메인 범위에서 크롤링 되어 일부 사회적 문제에만 초점을 맞춘 평가가 될 수 있음
  - 도메인 중복으로 인한 공정하지 않은 평가
- 저자는 익명의 작업자가 혐오 표현 데이터를 생성하도록 하는 것이 적절하다고 가정
  - 작업자가 task manager로부터 제공받은 prompt를 참고해 데이터를 생성 (최소한의 지침)
  - 크라우드 소싱 플랫폼을 이용해 작업자가 익명으로 데이터를 만들 수 있도록 함 (작업자의 불명예를 방지)
- System
  - crowd, task manager
  - pseudo 혐오 표현 분류기 구축 및 모델 배포
  - user-generated 데이터 수집 및 피드백
  - 3명의 task manager의 post-labeling

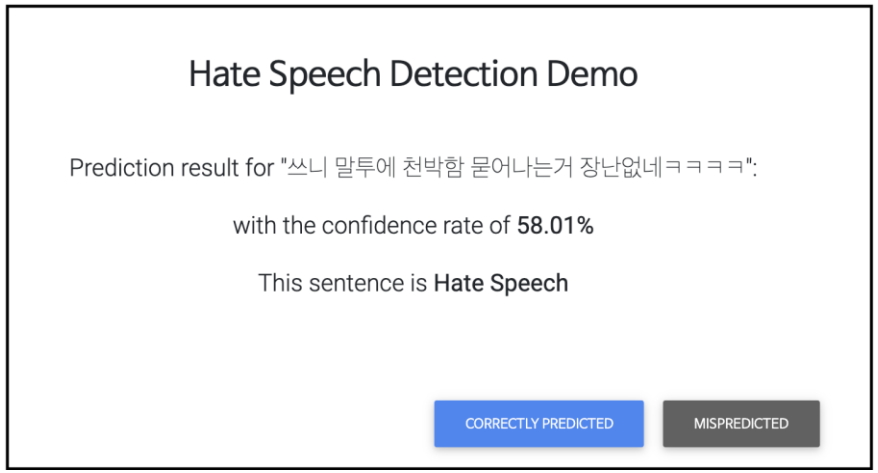


Introduction

- 기여
  - 기존의 데이터 구축 방식과 달리 작업자가 직접 혐오 표현 데이터를 생성하는 방법 제안
  - 새로운 한국어 혐오 탐지 평가 셋 구축 및 공개 (약 3K)
    - 라이선스 및 개인정보 문제가 없어 말뭉치 도메인의 잠재적 편향을 방지
  - 기존 라벨링 방식으로 구축된 데이터셋(벤치마크 데이터셋)과 성능 기반 비교를 통해 APEACH의 일반화 가능성이 특정 사전학습 말뭉치와 중복이 적다는 점에서 암시됨을 보임
    - 즉 기존 라벨링 방식으로 구축된 데이터셋에 비해 사전학습 데이터셋과 도메인 중복이 적음으로써 도메인 일반화 가능성을 보임 (도메인 중복으로 인한 성능 향상과 같은 문제를 방지)

System

- System 구성
  - pseudo 분류기 구축
  - 작업자 생성 데이터 수집 및 pseudo 분류기를 이용한 작업자로부터의 피드백
  - post-labeling
- pseudo 분류기 구축
  - 반복되는 작업으로부터 오는 집중력 저하를 방지하고 수집 과정에 적극적인 참여를 위해 pseudo 분류기를 사용함
  - 단순히 욕설 용어 사전을 만들고 이를 통해 pseudo-labeled 웹 텍스트 데이터셋을 만듦. 이를 단순 이진 분류기 학습에 사용
  - 분류기는 작업자가 생성한 text를 입력하면 텍스트가 편견이나 toxic 포함 여부 예측
- 분류기 배포 및 텍스트 수집
  - **배포**: 분류기는 작업자의 입력을 수집하고
    작업자로부터 피드백을 받기 위해 배포됨
  - **피드백**: 작업자가 생성한 텍스트를 입력하면 예측 라벨이
    추론되고 작업자의 의도와 일치 하는지 아닌지를 작업자가 선택
    이를 통해 ground truth 설정



Dataset

- 이와 같은 시스템을 이용해 작업자 생성 데이터셋 구축 (평가 데이터셋 APEACH)
- 작업자가 기초부터 혐오 표현을 생성하는 것은 어려울 수 있기 때문에 주제 제공
- Topic (10개)

- 작업자는 이 중 하나를 선택한 후 그에 따른 입력을 생성
- 주제는 랜덤한 순서로 제공됨 (잠재된 편향을 막기 위해)

Dataset

- Label
  - 문장이 혐오를 표현하는지 아닌지를 라벨로 정의함
  - 작업자는 혐오가 포함된 문장을 만들어야 하는지 혹은 혐오가 포함되지 않은 문장을 만들어야 하는지에 대해 지침을 받는다.
  - 혐오가 포함되지 않는 문장의 경우 주제는 동일하지만 공격적이지 않은, 중립적이거나 논란의 여지가 있어 보이지만 혐오는 아닌 발화를 나타내야함
  - ex) 나는 BLM 운동을 비하하는 사람이 싫다
    - 이는 '인종에 따른 차별 발언'을 주제로 하지만 혐오가 아님
  - 해당 연구에선 혐오 발언과 비혐오적 데이터가 균형을 이루어 제공됨

Dataset

- 세가지 요소(할당된 라벨, 모델의 예측 라벨, 작업자의 피드백)를 통해 최종 라벨 설정
- 예시
  - 할당 라벨=hate speech & 모델 예측=non-hate speech -> 작업자가 오분류로 체킹 -> ground truth는 hate speech로 저장됨

- 의심 데이터 제거
  - 이 과정에서 할당 라벨과 ground truth가 다른 데이터는 작업자의 실수로 간주되고 자동으로 제거됨
  - 즉 모델이 예측을 잘 했는데 작업자가 오분류로 체킹한 경우

- post-labeling
  - 데이터 생성 과정에서 다양한 윤리적 기준에서 비롯된 몇 의심스러운 데이터를 확인
  - 하지만 작업자 생성 데이터의 특성을 보장하기 위해 이러한 케이스의 데이터는 최소한의 task manager에 의해 확인됨
  - 이는 최종 결정에서 기존의 라벨링 및 투표 과정을 적용해 데이터의 퀄리티를 보장함
  - 세명의 task manager가 각 데이터에 대해 작업자의 피드백이 적절한지 확인함
    - task manager 모두 적절하지 않다고 판단한 데이터만 삭제함

Dataset

Dataset collection

- Compensation through moderator
  - 혐오 발언 데이터 생성에 대한 보상을 위해서는 근로자를 식별해야 하지만 수집 단계에서 익명성을 해칠 수 있기 때문에 크라우드소싱 플랫폼이 task manager와 작업자 사이 moderator(중재자) 역할을 할 수 있도록 함
  - 중재자만이 작업자의 프로필을 관리하는 방식으로 프로젝트가 설계됨
  - 작업자의 익명성 보장
- Worker selection for dataset quality
  - 모든 참여 희망 작업자가 해당 작업에 적절하진 않을 수 있음
  - 저품질 생성을 방지하기 위해 tutorial 과정이 존재
    - 작업자로부터 10개의 문장을 입력 받음
    - 잘못 라벨링 된 데이터의 비율을 계산 -> 이로부터 자주 실수를 만드는 작업자 제외
    - 가이드라인 예시와 입력이 동일한 경우
    - 입력이 한 글자 이하인 경우
  - 230명의 작업자 중 154명이 최종적으로 작업에 참여하도록 승인 됨

Dataset

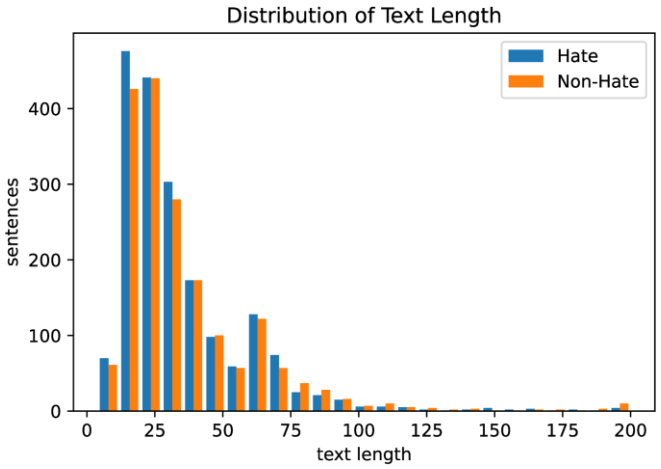
- Diversity of crowd-generated hate speech
  - 기존 익명 데이터 수집과 달리, 제안하는 방식은 주제 선택지를 사용해 다양한 주제의 텍스트를 생성하도록 요청할 수 있음
  - 작업자의 텍스트 생성을 가이드하고 이전의 혐오 발언 연구에서 덜 고려되었던 비혐오 발언 데이터를 수집하도록 함
  - heavy worker로부터의 데이터가 편향되는 것을 막기 위해 작업자 당 텍스트 생성을 최대 40개로 제한함

- Dataset summary
  - 해당 데이터셋 구축 scheme은 크라우드 소싱 플랫폼의 moderator를 통해 데이터 퀄리티, 주제 다양성, 데이터의 윤리적 문제를 보장할 수 있다.
  - 메인 단계에서 task manager 간 의견 불일치는 반대 클래스로 레이블 지정됨

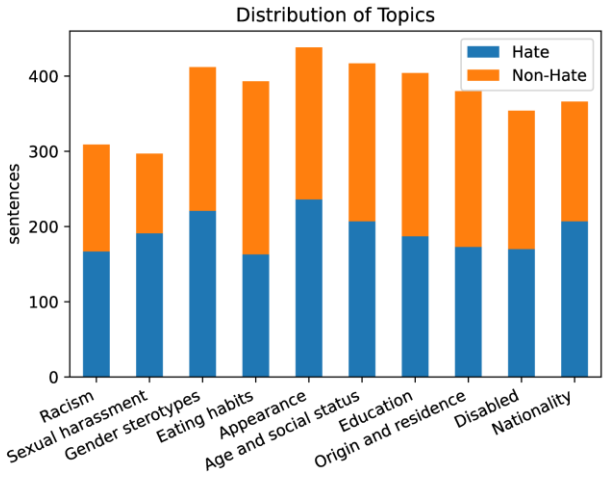
	Tutorial session		Main session	
	Non-hate	**Hate**	**Non-hate**	**Hate**
Accept	453	478	1386	1499
Reject	38	52	116	1
Total	491	530	1502	1500

  - 길이 분포
    - hate speech와 non-hate speech의 길이 분포가 유사
    - 이로써 암시적인 길이 편향을 방지
  - 주제 분포
    - topic prompt를 랜덤한 순서로 제공함으로써 습관적으로 상위 후보를 선택하는 경향에서 오는 편향을 방지

Dataset



<길이 분포>



<주제 분포>

- 혐오 발언 탐지 모델 평가
  - 한국어 혐오 발언 벤치마크 데이터셋인 BEEP! 데이터로 학습된 모델을 APEACH를 사용해 평가
  - 또한 BEEP!의 dev set과 APEACH를 평가 corpus로써 비교를 통해 일반화 가능성과 성능 경향을 확인

Experiment

- Korean Pretrained Language Models (PLM)
  - 공개된 한국어 사전학습 언어 모델을 사용

  - KoBERT
    - BERT 학습 scheme을 따르는 PLM (**한국어 위키피디아** 데이터로 사전 학습)
  - DistilKoBERT
    - DistilBERT의 distillation 기법을 이용한 KoBERT의 경량화 버전
  - KoELECTRA
    - 국립국어원이 발표한 **모두의 말뭉치, 한국어 위키피디아, 나무위키, 뉴스 기사** 등의 corpus로 사전 학습된 PLM
  - KcBERT
    - 12GB의 **네이버 정치 뉴스 댓글**로 사전 학습된 한국어 BERT 모델
  - SoongsilBERT
    - KcBERT에서 사용된 **정치 뉴스 댓글**에 더해 **대학교 커뮤니티**와 **모두의 말뭉치 데이터**를 사용해 사전학습한 RoBERTa 기반 모델

Experiment

- 학습 데이터
  - 파인튜닝에 BEEP! 데이터를 사용
    - BEEP!: 한국어 연예 뉴스 댓글에 hate, offensive, none 삼진으로 annotation 된 한국어 혐오 탐지 벤치마크 데이터셋
  - BEEP! 데이터셋 구축 scheme이 APEACH와 다르지만 두 데이터가 모델을 평가하는 경향을 확인하고 싶어 두 데이터셋을 모두 활용함
  - 두 데이터 모두 이진으로 동일하게 변형
    - BEEP! 에서 hate와 offensive를 합쳐 이진 데이터('hate+offensive', 'none')로 변형

- 평가
  - BEEP!의 dev set과 APEACH 사용
  - f1 score 계산

Experiment

- 결과
  - BEEP!으로 학습된 모델이 APEECH 에서도 합리적인 성능을 보여주었다.
  - 이는 해당 연구의 데이터셋 생성 기준이 기존 연구와 유사함(aligned)을 의미

Model	BEEP! dev set	APEACH (ours)	Relative difference
KoBERT	0.8030	0.7885	-1.81%
DistillKoBERT	0.7570	0.7715	1.92%
KoELECTRA-V3	0.7920	0.8101	2.29%
KcBERT-Base	0.8088	0.8086	-0.02%
KcBERT-Large	**0.8295**	0.8116	-2.16%
SoongsilBERT-Base	0.8261	**0.8424**	1.97%
SoongsilBERT-Small	0.8149	0.8228	0.97%
Composition	Hate + Offensive : 311 None : 160 Total : 471	Hate : 1,922 Non-hate : 1,848 Total : 3770	

- 코퍼스 도메인의 영향
  - BEEP!의 경우 KcBERT-Large에서 크게 좋은 성능을 보임
  - APEACH에서 KoELECTRA(BEEP! dev set에선 낮은 성능)가 KcBERT와 유사한 성능을 냄
  - 이는 사전학습에 사용된 데이터의 도메인과 스타일이 다운스트림 태스크 성능 측정에 영향을 미친다는 것을 의미함

Experiment

- 주제별 성능
  - 주제별 정확도 확인
  - 편차는 학습 셋과 평가 셋의 구성 방식 차이에서 비롯된 것으로 보임
  - BEEP!에는 Gender stereotypes과 Sexual harassment를 다수 포함하는데 topic별 성능에서는 낮게 나옴
  - 명확한 이유는 알 수 없으나 학습셋과 평가셋 간에 텍스트 스타일 불일치가 원인으로 추측됨

Topic	F1 Score
Nationality	0.8519
Age and social status	0.8700
Eating habits	0.8182
Appearance	0.8114
Gender stereotypes	0.7993
Sexual harassment	0.7610
Racism	0.8511
Origin and residence	0.8393
Disabled	0.8525
Education	0.9035

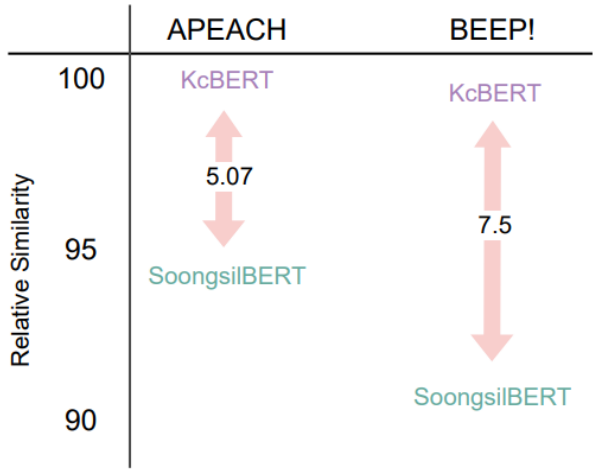
Table 2: SoongsilBERT-Base's F1 score of binary classification according to topics.

Analysis

- 도메인 일반화 가능성
  - APEACH는 도메인 의존성 문제를 해결
    - prompt에 기반해 작업자가 혐오 발언을 생성하도록 함
    - 생성할 텍스트의 스타일을 지정하지 않음
    - 이 둘은 기존 annotation scheme에서 보장되지 않음 -> 이는 BEEP!의 단점
    - BEEP!은 뉴스 댓글이지만 사전 학습 과정에서 뉴스 댓글이 포함되지 않은 KoELECTRA에서 상대적으로 낮은 성능을 보여줌. 반면에 APEACH에서는 다른 경향을 보임
    - APEACH는 모델 사전학습에 사용된 corpus의 도메인에 의존성을 덜 갖고 성능을 평가할 수 있다.
  - SoongsilBERT
    - BEEP!과 APEACH에 대한 경향이 다름
    - KcBERT에서 BEEP이 도메인 특화로 인해 이득을 봄. 즉 더 좋은 성능을 냄
    - BEEP 에서 KcBERT의 좋은 성능은 기존 annotation scheme이 특정 도메인에 의존성을 가져다 주었기 때문일 수 있음
    - 이는 평가에서 도메인 일반화 가능성의 한계로 작용함

Analysis

- 도메인 일반화 가능성에 대한 분석
  - 각 평가셋과 PLM 사전학습 데이터 사이 TF-IDF 유사도를 계산함으로써 각 평가셋의 도메인 일반화 가능성을 검증함
  - 두 개의 사전학습 말뭉치와 두 개의 평가 데이터셋을 사용해 총 4개의 score를 계산했고, 이를 최대값으로 정규화
  - TF-IDF 단어 사전은 4개의 코퍼스의 whitespaced 단어를 기반으로 구축
  - 평가셋과의 코사인 유사도 계산의 타당성을 위해 KcBERT와 SoongsilBERT 사전학습 데이터의 1%를 무작위 샘플링
  - 결과
    - BEEP!에서는 두 사전학습 코퍼스 사이 뚜렷한 갭이 있지만 반면에 APEACH는 상대적으로 작은 갭을 보여준다
    - 이는 APEACH가 도메인 중복에 덜 민감함을 의미하며 평가의 일반화 가능성을 시사함



< 유사도 계산 결과 >

Analysis

- 정성 분석
  - APEACH에서 BEEP!에서 가장 많이 관찰된 'Gender stereotypes'와 'Sexual harassment'에 대해 조사함
  - 첫 두 개의 text는 웹 댓글 스타일(BEEP!과 유사)이며 나머지는 일반적인 스타일의 text
    1. "ㅅㅂ옷 꼬라지 하고는.. 게이냐?"
    2. "여자 연예인들은 다들 뒤로 스폰끼고 구멍 장사하는거 아닌가?"
    3. "아무래도 아이를 돌보거나 그 밖의 집안일 에 관해서는 남성들보다는 여성들이 좀 더 신경 써야 하는 것이 당연하다고 생각합니다."
    4. "남자면 그냥 해도 될텐데 고집을 많이 부리네요. 불편해도 좀 참아야 다들 편한데"

  - BEEP!으로 파인튜닝된 숭실버트가 1,2번은 제대로 toxicity를 추론했으나, 3,4번 추론엔 실패함
  - 정중한 매너로 고정관념을 내포한 댓글인 3,4번을 제대로 추론 못함
  - 즉 학습된 데이터셋과 유사한 댓글 스타일에 대해서는 추론을 제대로 했으나, 스타일이 다른 텍스트에 대해선 그러지 못함

Analysis

- APEACH로 학습
  - APEACH로 학습한 모델의 낮은 성능
  - 추정 원인
    - 훈련 데이터셋의 크기 차이
      - BEEP-8K / APEACH-3.7K
    - BEEP!과 composition(style 등)이 다름

	Validation	
Train	**APEACH**	**BEEP! dev**
APEACH	-	KoELECTRA: 0.7502 KcBERT-Large: 0.7893
BEEP! train	KoELECTRA: 0.8101 KcBERT-Large: 0.8116	KoELECTRA: 0.7916 KcBERT-Large: 0.8295

Conclusion

- 결론
  - 크롤링 및 라벨링 기반의 기존 데이터셋 구축 방식과 달리 작업자 기반 데이터 구축 방식을 소개
  - 작업자의 익명성과 데이터의 신뢰성을 모두 보장하는 작업자 기반 데이터셋 구축 이후, 이전 연구와의 비교를 통해 분석을 제시
  - APEACH는 작업자 기반 데이터 구축 기법이 도메인 일반화(domain generalizability)와 주제 다양성(topic variety)을 이룸

추가

- Train-Test 중복 완화
  - BEEP!이 연예 섹션의 뉴스 댓글을 다루고, KcBERT가 정치 뉴스 댓글을 기반으로 사전학습 되었다는 점에서 유사한 도메인을 공유
  - 그러므로 잠재적인 토큰 중복이 존재할 것
  - 그렇다면 KcBERT에서 BEEP! 평가셋을 사용했을 때 좋은 성능을 낼 것이다. 하지만 이는 적절치 않음

  - 이를 방지하기 위해 도메인 중복을 고려한 APEACH를 제안
  - 그렇기 때문에 APEACH가 더 넓은 범위의 말뭉치로 사전학습 된 PLM(soongsil BERT) 평가에 더 적합할 것으로 생각
  - 또한 기존 annotation scheme에서 train과 test 셋 중복의 위험을 강조하고 APEACH가 이러한 점을 완화할 수 있다는 것을 강조
  - 이러한 점은 APEACH의 훈련 셋으로서의 유용성을 보장함