HateScore 인공지능 한국어 혐오발언 분석기술

발명도둑잡기

https://github.com/sgunderscore/hatescore-korean-hate-speech

https://github.com/sgunderscore/hatescore-korean-hate-speech/blob/main/rsc/zoomed_HateScore_transparent.png" style="background-color:transparent;">

HateScore : Human-in-the-Loop and Neutral Korean Multi-label Online Hate Speech Dataset (feat. SmilegateAI UnSmile Dataset)

- 본 깃헙 페이지는 https://arxiv.org/abs/2204.03262" style="background-color:transparent;">(Kang, Tayoung, et al., 2022) 논문에서 활용한 데이터셋 중 보조 데이터셋 1.1만건을 다룹니다. 논문의 메인 데이터셋인 Korean UnSmile Dataset은 Smilegate AI의 기획과 지원을 바탕으로 개발되었습니다.
- 본 데이터셋의 크기는 약 1.1만 건으로, Korean UnSmile Dataset의 base model을 활용해 HITL(Human-in-the-Loop) 방식으로 태깅된 1.7천 건, 위키피디아에서 수집한 혐오 이슈 관련 중립 문장 2.2천 건, 규칙 기반으로 생성된 중립 문장 7.1천 건의 세 가지로 구성되며 중립 문장 오분류 방지를 주 목적으로 개발되었습니다.
- 언더스코어는 Korean UnSmile Dataset의 개발과 레이블링 작업을 진행했으며, 본 HateScore 데이터셋 역시 당시의 참여 인원 및 레이블링 기준을 동일하게 유지했습니다. 다만 이하의 '4.권장사항' 및 '5.FAQ' 항목은 SmilegateAI의 공식 입장과는 별개로 언더스코어가 작성한 독립적 의견입니다.
- 데이터 수집 및 레이블링 방식, 혐오발언 유형 선정 기준 등 보다 상세한 정보는 https://arxiv.org/abs/2204.03262" style="background-color:transparent;">논문에서 확인 가능하며, 국문 요약은 여기서 보실 수 있습니다.
- Huggingface Demo에서도 직접 문장 분류 테스트를 해보실 수 있습니다.
- 본 데이터셋은 비영리·학술 목적의 활용을 전제로 공개되었습니다.

https://github.com/sgunderscore/hatescore-korean-hate-speech#1-%EC%A0%81%EC%9A%A9-%EC%98%88%EC%A0%9C-kcbert-base-unsmilehatescore" style="background-color:transparent;float:left;padding-right:4px;margin-left:-20px;line-height:1;">1. 적용 예제 (KcBERT-base, Unsmile+HateScore)

문장	여성	성소수자	남성	인종	지역	종교	연령
여자는 집에서 애나 봐라	0.86	0.01	0.03	0.03	0.01	0.01	0.01
조ㅈ족은 21세기의 홍어다	0.03	0.02	0.03	0.68	0.89	0.04	0.03
너는 전라도 사람이니?	0.00	0.00	0.00	0.00	0.01	0.00	0.00
상폐 한남들 다 재기하라고	0.09	0.02	0.88	0.05	0.05	0.04	0.55
도심에서 변태성욕 축제라니 말세	0.06	0.79	0.02	0.01	0.13	0.01	0.01
쉰내나는 태극기들 틀니 압수	0.07	0.03	0.06	0.09	0.03	0.05	0.94
개독이나 짱_깨나 거기서 거기	0.05	0.03	0.02	0.84	0.09	0.92	0.05
저 친구는 필리핀 출신이다	0.00	0.00	0.00	0.01	0.00	0.00	0.00
쿵쾅이들도 필리핀 그지는 싫지?	0.74	0.01	0.04	0.71	0.02	0.01	0.01

```
>>> from transformers import TextClassificationPipeline, BertForSequenceClassification, AutoTokenizer
>>> model_name = 'sgunderscore/hatescore-korean-hate-speech'
>>> model = BertForSequenceClassification.from_pretrained(model_name)
>>> tokenizer = AutoTokenizer.from_pretrained(model_name)
>>> pipe = TextClassificationPipeline(
        model = model,
        tokenizer = tokenizer,
        device = -1, # gpu: 0
        return_all_scores = True,
        function_to_apply = 'sigmoid')
>>> for result in pipe("착한 중국인은 죽은 중국인이다")[0]:
        print(result)

{'label': 'None', 'score': 0.07771512866020203}
```

```
{'label': '기타 혐오', 'score': 0.02803093008697033}
{'label': '남성', 'score': 0.013538877479732037}
{'label': '단순 악플', 'score': 0.01559345331043005}
{'label': '성소수자', 'score': 0.014305355027318}
{'label': '여성/가족', 'score': 0.014650419354438782}
{'label': '연령', 'score': 0.014001855626702309}
{'label': '인종/국적', 'score': 0.9227811098098755}
{'label': '종교', 'score': 0.035127196460962296}
{'label': '지역', 'score': 0.02069076895713806}
```

https://github.com/sgunderscore/hatescore-korean-hate-speech#2-%EB%8D%B0%EC%9D%B4%ED%84%B0%EC%85%8B-%EB%B9%84%EA%B5%90" style="background-color:transparent;float:left;padding-right:4px;margin-left:-20px;line-height:1;">2. 데이터셋 비교

Model Performance : LRAP (Label Ranking Average Precision)

모델명	Unsmile	Unsmile+HateScore
KcBERT-base	.886	.914
KcBERT-large	.892	.919
KcELECTRA-large	.884	.912

Base Model 기준 비교 예제 (표 안의 값은 혐오발언 분류 확률)

혐오발언 분류 확률	Unsmile	Unsmile+HateScore
저 사람 중국인이네	0.87	0.20
너 페미니스트니?	0.03	0.01
동성혼은 논쟁적이지	0.35	0.01
무슬림을 다 죽인다고?*	0.84	0.76

*두 모델 모두 오분류한 사례

https://github.com/sgunderscore/hatescore-korean-hate-speech#3-%EC%9D%B8%EC%9A%A9-%EB%B0%A9%EC%8B%9D" style="background-color:transparent;float:left;padding-right:4px;margin-left:-20px;line-height:1;">3. 인용 방식

논문

```
@misc{https://doi.org/10.48550/arxiv.2204.03262,
  doi = {10.48550/ARXIV.2204.03262},
  url = {https://arxiv.org/abs/2204.03262},
  author = {Kang, TaeYoung and Kwon, Eunrang and Lee, Junbum and Nam, Youngeun and Song, Junmo and Suh, J
  keywords = {Computation and Language (cs.CL), Computers and Society (cs.CY), FOS: Computer and informat
  title = {Korean Online Hate Speech Dataset for Multilabel Classification: How Can Social Science Improv
  publisher = {arXiv},
  year = {2022},
  copyright = {arXiv.org perpetual, non-exclusive license}
}
```

https://github.com/sgunderscore/hatescore-korean-hate-speech#4-%EA%B6%8C%EC%9E%A5%EC%82%AC%ED%95%AD" style="background-color:transparent;float:left;padding-right:4px;margin-left:-20px;line-

height:1;">4. 권장사항

- HateScore는 중립 문장을 포함하고 2021년도 하반기 이후의 댓글 데이터 역시 포함한다는 강점이 있으나, 3인의 다수결 투표로 최종 레이블을 결정한 UnSmile과 달리 Human-in-the-loop 방식으로 '모델의 분류 확률'과 '연구원 한 명의 의견'의 두 가지 값만을 활용했습니다. 이에 응용 시에는 HateScore와 UnSmile 데이터를 함께 학습하는 것이 좋습니다. https://arxiv.org/abs/2204.03262" style="background-color:transparent;">논문 역시 이와 동일한 방식을 사용했습니다.
- HateScore는 온라인 '댓글' 데이터만을 다룹니다. 그렇기에 "아까 학력 인증한 연베대게이다. 학점 ㅁㅌㅊ?"나 "페미니스트들의 실체.png"와 같은 웹 커뮤니티 제목 텍스트에 모델을 적용할 경우, 혐오발언 여부를 오분류할 가능성이 높습니다. 이에 댓글 텍스트에만 적용하는 것을 권장합니다.
- 각 혐오발언 카테고리를 독립적으로 간주하는 단순 분류기 대신 멀티레이블(multi-label) 방식의 분류기 개발을 권장합니다.
- 입력한 텍스트가 혐오발언 카테고리에 해당되지 않더라도 '단순 악플'에는 해당될 수 있으니, 멀티레이블 분류기에서 주어진 댓글의 공격성을 단순히 "1-(Clean 분류 확률)"만으로 계산하는 것은 부적합합니다.

https://github.com/sgunderscore/hatescore-korean-hate-speech#5-faq" style="background-color:transparent;float:left;padding-right:4px;margin-left:-20px;line-height:1;">5. FAQ

- 혐오발언 유형은 어떻게 되나요?
  → 여성, 성소수자, 지역, 인종/국적, 종교, 연령, 남성의 7가지 이며 기타 혐오발언, 단순 악플, 일반 댓글(clean)의 3가지 유형이 추가로 제공됩니다.
- 혐오발언 카테고리 별 데이터 수는 중요도와 비례하나요?
  → 아니요. 그렇지 않습니다.
- 기타 혐오발언의 경우 어떤 내용을 포함하나요?
  → 외모에 대한 조롱, 특정 직업군에 대한 비하, 장애 희화화 등 위 7가지 대분류에 포함되지 않는 혐오발언들이 이에 해당됩니다.
- 기타 혐오발언은 7가지 유형보다 중요하지 않은가요?
  → 아니요. 그렇지 않습니다. 예산 및 시간의 제약으로 인해 우리 사회에 존재하는 모든 유형의 혐오발언에 대해 충분한 수의 데이터셋을 개발할 수는 없었습니다.
- 왜 '남성'이 혐오발언의 유형 중 한 가지에 포함되어 있나요?
  → 물론 좁은 의미에서의 혐오발언은 사회적 소수자(social minority)에 대한 적대적 발언만을 지칭합니다. 다만 그럼에도 현실에서 규모가 빠르게 성장 중인 특정 유형의 악플을 아예 무시하는 역시 바람직하지는 않기에, '온라인' 악플·혐오발언을 주제로 하는 본 데이터셋의 특성상 남성 카테고리를 포함시켰습니다. 즉, 실제 현실에서 여성에 대한 경제적 차별, 문화적 고정관념, 범죄 피해 관련 문제가 여전히 존재하는 것과는 다소 독립적인 논의라고 판단했습니다. 이와 비슷하게 '종교' 카테고리에서도 국내 종교인구 1위를 기록하고 있는 개신교에 관한 적대적 발언들 역시 혐오발언으로 분류했습니다. 보다 상세한 논의는 앞서 소개한 국문 논문 요약 및 https://arxiv.org/abs/2204.03262" style="background-color:transparent;">논문 본문에 작성되어 있습니다.
- 왜 '운지'나 '재기'와 같은 표현들은 혐오발언이 아닌 단순 악플로 분류되었나요?
  → 모든 온라인 은어가 그러하듯 시간이 흐르면 초기의 의도와는 다른 방식으로 활용되고는 합니다. '운지'와 '재기'의 경우, 각각 노무현 전 대통령과 성재기 전 대표의 투신에 대한 조롱으로 시작되었으나 일정한 시간이 지난 현재에는 일베 및 워마드에서 하락, 감소, 추락, 죽음 등을 지칭하는 보다 포괄적인 은어로 활용됩니다. 가령 "왜 나스닥 운지하고 있노이기?"나 "요새 워녬글 수 재기했노;;"와 같은 표현들이 사용될 때마다 발화자들이 매번 노무현과 성재기 두 명의 인물에 대한 공격을 의도하지는 않습니다. 물론 "홍어 틀딱들 코로나로 빨리들 운지했으면"이나 "젠신병자들 정병 걸려서 재기했으면 좋겠노"와 같이 여타 혐오 표현과 함께 활용될 경우에는 혐오발언으로 태깅하였습니다.