1. 연구 배경

포털 사이트, SNS, 커뮤니티(에브리타임 등) 같은 온라인 플랫폼에서 사이버 언어 폭력이 최근 점점 증가하고 있다. 온라인상의 혐오 표현은 다양한 형태로 서비스 이용자들에게 피해를 끼치고 있다. 폭발적인 댓글의 양으로 인해 이를 인간이 하나하나 관리감독하기란 어려운 실정이다.

2. 연구 목적

- 댓글 상의 혐오 표현을 감지하는 LSTM-Attention 기반의 딥러닝 모델을 구현하는 것입니다.  
- 학습을 위해 스마일 게이트의 언스마일 데이터셋, korpora의 한국어 혐오 데이터셋을 사용하고 이 데이터로 학습된 모델을 한양대 에브리타임 커뮤니티에 적용하여 성능을 평가하는 것입니다.

- 예측 시점의 Attention score를 분석해서 어떤 단어들이 모델의 의사결정에 많은 영향력을 끼쳤는지 파악하고자 합니다

- 나아가, 이를 활용해 커뮤니티 관리자가 제재하기 쉽게 데이터를 가독성 있게 정리해서 관리자에게 전해주고자 합니다.

- 시간적 여유가 있다면 다른 아키텍처를 사용해서 개발하고자 합니다. 그렇게 accuracy를 비교하고자 합니다.

3. 데이터 구성

- 학습 데이터: 언스마일 데이터셋(스마일게이트AI), 한국어 혐오 데이터셋(Korpora)

- evaluation: 한양대 에브리타임으로부터 크롤링 한 후 라벨링

4. 설계 및 구현

공부해서 하겠슴다..

5. 결과 및 증명

공부해서 하겠슴다.