
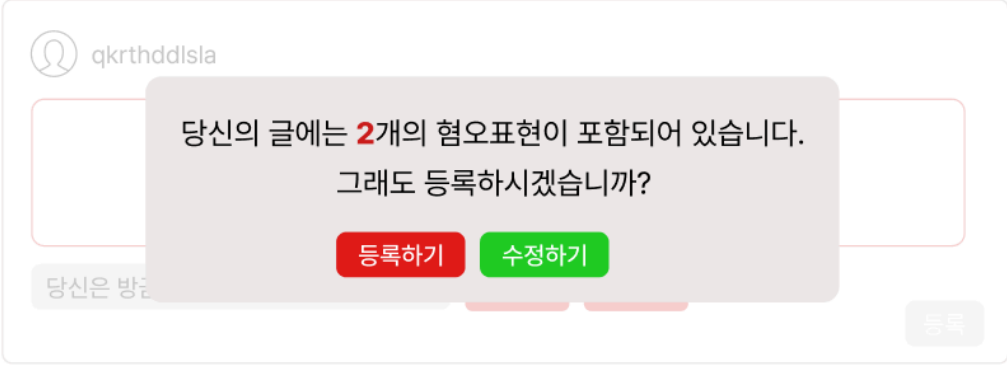


PoC Lab - 아이펠톤 프로젝트 계획서

개발아이템명	korean hate speech detection AI		
소속	아이펠 리서치 파트		
신청자 성명	박송이	담당퍼실	근철님

□ 프로젝트 아이템 개요(요약)

아이템 소개	<p>혐오표현 탐지 모델</p> <p>구체적 혐오 표현을 단어수준으로 탐지해서 알려준다.</p>
아이템의 특징 및 차별성	<ul style="list-style-type: none"> - 대형 포털이 자사 서비스를 위한 탐지기가 아닌 소기업들도 사용할 수 있는 API 서비스 제공 - (코딩을 알지 못해도 쉽게 자사 서비스에 적용 가능) - 구체적 혐오표현 단어를 탐지해 제공 - 뉴스 댓글뿐만 아니라 많은 여론이 모이는 커뮤니티, 유튜브 댓글까지도 활용해 다양한 범위의 혐오 표현 탐지
이미지	 <p>Frame 2</p> 

1. 문제인식 (Problem)

1-1 프로젝트의 목표 및 목적(필요성)

● If not now, when? 지금이 아니라면 도대체 언제 할 것인가?

- 악성댓글에서 혐오댓글로 이젠 혐오사회... 지금 무엇을 하지 않는다면 그 다음은 무엇이 올 것인가?
- 왜 소수의 사람들, 그것도 누군가를 혐오하는 목소리가 과잉 대표화되어 있는가. 이젠 악순환을 막자.

24 아이뉴스24 | 2004.02.17. | 네이버뉴스

[댓글문화를 바꾸자-하]"시간걸려도 자율정화가 최선"

MBN | 2007.01.23. | 네이버뉴스

"악성댓글 범죄 5년새 2.5배 증가"

3기 뉴스 쿠키뉴스 | 2006.01.25. | 네이버뉴스

"전체 네티즌의 0.06%인 악플러가 25%의 악플 양산"

● 혐오사회로의 악순환 방지

- 악성 댓글로 연예인이 사망하던 문제에서 이제는 한 개인이 악성댓글과 혐오표현으로 세상을 등지고 있음.
- 연예인을 향한 악성 댓글로 연예 뉴스기사에서 더 이상 댓글을 달 수 없게 되었는데 이제는 그 혐오가 확산되어 사회재난 기사에도 댓글을 폐지하자는 의견이 나오고 있음.

OSEN | 2007.01.21. | 네이버뉴스

유니 자살로 다시 도마 위에 오른 '악플' 파문

“2차 피해 막자” 사회재난 기사 댓글 폐지가 답?

YTN PICK | 2022.12.14. | 네이버뉴스

이태원 참사 생존자 추정 10대 숨진 채 발견

● 혐오표현으로 인한 사회적 손실 비용 감소

- 혐오발언(hate speech)은 단순히 거시적인 사회 이슈일 뿐만 아니라 사람이 모여서 수익을 내는 IT 플랫폼 기업에게도 중요한 문제
- 혐오표현으로 인한 전체 사회적 비용은 앞으로도 증가할 추세

악성 댓글 사회적 비용 年 최대 35조... “피해자 구제 개선 필요”

(단위:억)	최소	최대
불안·우울에 따른 행복 상실	28조 9335억	
스트레스로 인한 능력 저하	1조 4095억	2조 8189억
변호사 선임과 손해배상	1433억	3조 5229억
합계	30조 5371억	35조 3480억

자료:바른ICT 연구소

1-2 아이템의 독창성

● 기존 프로젝트의 한계와 장점

1. 스마일 게이트 - 혐오에 대한 양적 연구를 시도하였지만 데이터셋의 크기는 작음(30,000)
2. 튜입 서비스 - 별도의 코딩지식 없이도 편하게 사용 가능, 윤리성 판별 기능 제공
3. 네이버 시클린봇 : 데이터셋의 크기는 적정하나 네이버 뉴스 데이터만을 가지고 혐오표현 학습, 2(혐오), 1(애매함), 0(혐오없음)으로 비교적 간략하게 태깅함.

● 이번에 개발하고자 하는 서비스의 차별점

1. 미시적 변화부터 거시적 맥락까지 인간이 혐오하는 방식을 학습한 인공지능 모델 개발
2. 혐오표현 감지를 하는 ai bot을 피하기 위해 사용한 오탈자의 패턴도 학습한 ai model 개발
3. 풍부한 표현 학습으로 혐오표현 있음/없음 이상으로 다양한 혐오표현의 패턴을 감지(구체적 혐오의 종류 ex) 성차별, 지역 차별, 인종차별 등등)
4. 데이터 보강을 통한 맥락적 이해 시도 (pragmatic understanding)
 - 2012년 이후부터 데이터 수집 진행 (맥락적 이해 시도)
 - 제목에 있는 텍스트도 활용 (뉴스의 제목과 댓글의 맥락을 함께 파악 시도)
 - 유튜브 사이버 렉카 채널들 댓글 이용

2. 개발 및 연구 내용

2-1. 구현 내용 상세(구현 가능성)

혐오표현 탐지 방법 3가지

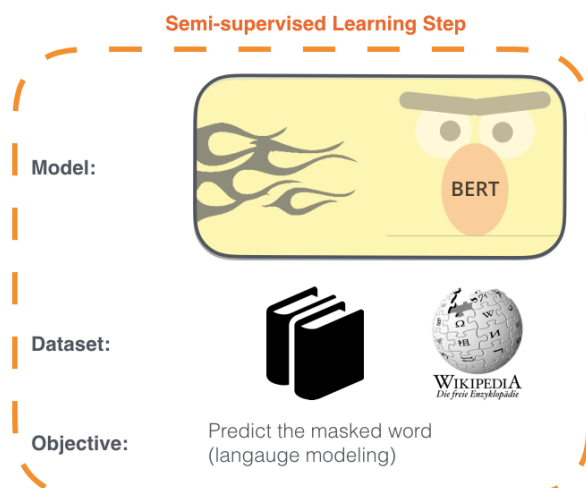
1. 키워드 기반 분류 : 특정 키워드를 기반으로 혐오 표현 탐지 but 중의적인 표현이나 뉘앙스 파악에는 한계점이 있음.
2. 인간코더 기반 분류(사람이 직접 라벨링) -> 인간이 라벨링한 자료를 바탕으로 학습을 시키는 방식 -> 고비용, 주관성 포함의 위험성
3. 커뮤니티 기반 분류 -> 온라인 커뮤니티로부터 추출된 자료로 학습 -> 커뮤니티 내에서만 혐오 표현을 학습해서 일반화에 한계 있음.

●설계의 방향성

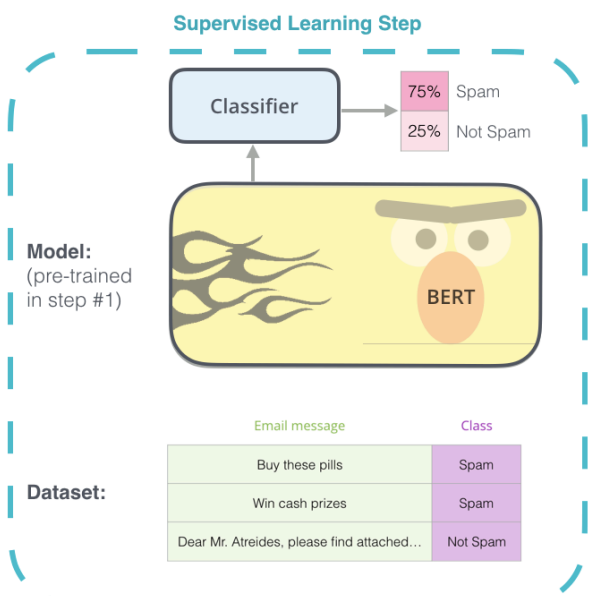
1. semi-supervised training : 뉴스 댓글창, 커뮤니티 댓글, 커뮤니티 글 제목 등 다양한 댓글이 존재하는 커뮤니티로부터 댓글을 크롤링 하여 학습 진행
2. supervised-training : 아래에 있는 혐오 표현 데이터셋을 이용하여 학습 진행

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



2 - Supervised training on a specific task with a labeled dataset.

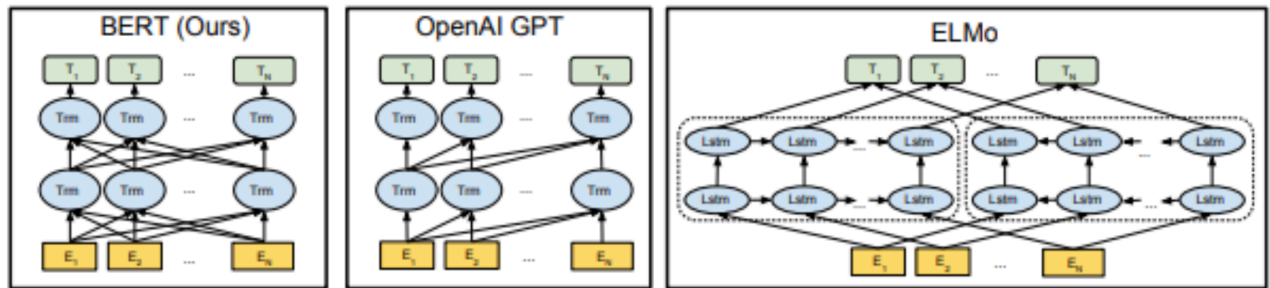


<출처 : <https://jalammar.github.io/illustrated-bert/>>

● Bert 모델의 차별성

ELMo는 좌-우(left-to-right), 우-좌(right-to-left) 문맥을 각각 독립적으로 계산하여 접합한 형태
OpenAI GPT는 좌-우(left-to-right)로만-우(left-to-right) 계산하는데 두 방법 모두 공통적으로 양방향 문맥 등을 보지 못하여 충분히 언어 표현을 하지 못하는 단점이 있음.

이에 비해 BERT는 사전 학습을 위해 두 가지 방법 (Masked Language Model(MLM)과 Next Sentence Prediction(NSP))를 사용하는데 이 방법들은 BERT가 양방향으로 학습되어 문맥을 더 잘 파악할 수 있게 함. <출처 : <https://happy-obok.tistory.com/23>>



2-1-1. 활용 가능한 데이터셋 정리

Korean Hatespeech Dataset	KoreanUnSmile Dataset	HateScore (unsmile의 보조)	korean-malicious-comment Dataset	curse-detection Dataset
labeled unlabeled news_title	labeled	데이터셋 1.1만건	비정제 데이터에 0,1로 라벨링	5,825문장을 분류
7,896 training set 471 validation set 974 test set	16,102 training set 3,985 validation	HITL - 1.7천 건, 위키피디아 혐오 문장 2.2천 건, 규칙 기반으로 생성된 중립 문장 7.1 천 건		
social bias (gender/other/none) hate speech (hate/offensive/none) 2,033,893 unlabeled	혐오표현(다중레이블) 악플/욕설		hate ->0 none ->1	hate ->1 none -> 0

2-1-2. PLM(Pretraining Language Model) 정리

1. BERT	2. ELECTRA	3. GPT	4. Encoder-Decoder	BART
KoBERT(SKT)	TUNiB-Electra(TUNiB)	LMKor-KoGPT2	KE-T5	KoBART
KorBERT(ETRI)	KoELECTRA	KoGPT2(SKT)	LMKor-T5	Korean-BART
HanBERT	KcELECTRA	KoGPT-Trinity 1.2B(SKT)	KcT5	
KcBERT	KoCharELECTRA	KoGPT(kakaobrain)	pko-t5	

KR-BERT (SNU)	LMKor-ELECTRA		KoT5(Wisenut)	
Ko-Char-BERT			ET5(ETRI)	
LMKor-BERT				
DistilKoBERT (SKT 경량화)				
KLUE-BERT				
KPFBERT				
brainsbert (kakaobrain)				
KalBert				
LMKor-Albert				
KLUE-RoBERTa				
KoBigBird				

2-2. 개발 아이템 기대효과

- 향후 연구로서의 가치, 사회적 합의가 필요한 AI 주제에 대한 깊이있는 기념비적 연구가 되리라 기대
- 인구감소 문제로 앞으로 이민에 대한 사회적 포용과 이해가 요구되고 있는 상황에서 한국도 단일민족 국가가 아닌 다민족 국가로 나아갈 확률이 큼. 이런 상황에서 문화와 종교, 언어가 다른 사람들이 한국 사회에서 함께 살아갈 때 극심한 혐오와 반대 정서가 생겨날 것으로 생각됨. 이미 다민족 국가로서 많은 이민자들이 함께 살아가고 있는 외국의 모델을 참고하여 다민족 국가로의 준비를 시작하고 그 때 지금 개발한 모델이 국가의 혼란과 혐오를 조금은 줄여주는 역할을 할 수 있기를 기대함.

‘빠~당신 댓글, 심해요’ 알고리즘으로 혐오 잡는 스웨덴

- 사기업 뿐만 아니라 공기업이나 정부 산하기관에서 민원인과 온라인으로 대화를 해야할 경우, 사람들이 혐오 발언을 사용할 수 있음. -> 대한민국 정부 민원 시스템에 혐오 표현 감지 등으로 더욱 확대하여 활용 가능
- ai 봇을 활용하여 혐오표현을 미리 필터링할 수 있다면 개인적으로 불필요한 감정낭비, 시간낭비 등을 막을 수 있고 사회적으로는 사회적 비용을 줄일 수 있음.
- 고객과의 비대면 소통 채널에서 고객이 혐오 발언을 남긴 경우 미리 필터링하여 감정노동자의 일터 환경 개선에 이바지할 수 있음.
- IT 플랫폼 기업 ex) 요기요, 배달의 민족 등 고객이 서비스를 이용하고 리뷰를 남기는 경우 악의적인 혐오 발언을 필터링 가능
- 최근 국회, 학회 등에서 악성댓글 뿐만 아니라 혐오댓글, 혐오사회에 대한 연구와 토론을 활발히 진행하고 있는만큼 이 주제에 사회적 관심사가 높고 이를 AI를 활용하여 적절히 대응할 수 있다면 학계, 정부, 기업 등에서 모두 유용하게 사용될 수 있으리라 기대됨.



- 또한 국회에서 조치를 취한 플랫폼 사업자에 대한 인센티브를 주는 방향도 고려되고 있는 만큼 단순히 사회를 위한 이로운 행위라는 점을 넘어서 사회에 환원하는 기업이라는 긍정적 이미지와 약간의 인센티브까지 얻을 수 있으리라 기대됨

국회서 현황-대응방안 토론회

악성댓글 부정적 영향 첫 계량화... '2명 중 1명' 직·간접 피해 경험

“플랫폼 사업자 조치 인센티브 줘야... 표현 자유 억압 않는 수준 ID 공개”

3. 실행 계획

3-1. 기간내 프로젝트 구현 완성을 위한 전략

1. 이해의 baseline 구축하기

- [\[뉴스기사\] 혐오는 어떻게 정당성을 얻나\[혐오의 민낯\]](#)
7명이 말하고 93명은 침묵하는 온라인 담론... 정치인과 언론이 확산하며 지배여론처럼 포장
- [\[특집기사\] 왜 혐오의 시대가 되었나?](#)
- [\[뉴스기사\] "혐오 댓글 걸러 드려요"](#)
- [\[테크리뷰\] 클린봇 2.0: 문맥을 이해하는 악성 댓글\(단문\) 탐지 AI](#)

2. 데이터셋 구축 (30%)

- 데이터 선정
- 크롤링
- 기본적인 데이터 전처리
- EDA
- 임베딩 방법 선정
- 토큰나이저 선택

2. PLM 활용 (20%)

- KcBERT, TUNiB-ELECTRA, KoELECTRA, KcELECTRA 등을 이용한 baseline 만들기
- baseline에서 모델 구조 일부 변경 실험 진행

3. 모델 설계 및 수정보완의 반복작업 (35%)

- 손실 함수 정의
- 모델 구조 설계
- 학습
- 결과를 확인하고 필요하다면 추가 데이터 정제나 다른 임베딩 방법 선택
- 위의 4가지 단계를 반복 실험
- 최종 모델 선택

4. 결과 정리 및 GitHub 페이지 제작과 작업 내용 공유 (10%)

- Github 페이지, velog 등을 통해 작업한 과정을 정리해서 공유

5. Presentation 준비 (5%)

- 데이터셋 구축 방법 -> 데이터 전처리 과정 -> 임베딩 선택 시 고려한 사항 공유
- 다른 PLM을 이용한 baseline 제작 결과 공유
- 모델 설계 시 어려웠던 점, 중점을 두었던 부분 공유
- 최종 모델 선택 시 고려한 사항과 모델이 가지는 한계점, 나아가야 할 방향 공유

3-2. 아이펠톤 기간 내 마일스톤

Task	비중	목표기간	세부내용
데이터셋 구축	30%	3주	기존에 있는 데이터셋에서 부족한 데이터들을 보완하고 전처리 하는 단계
PLM 활용	20%		미리 학습된 모델들을 가지고 데이터를 돌려보고 baseline을 만드는 단계
모델 설계 및 수정보완	35%	3주	baseline을 기반으로 직접 모델을 만들어 특정 단어가 포함되지 않아도 전체 맥락을 파악할 수 있는 모델 설계
결과 정리 및 공유	10%	1주	제작 과정을 웹페이지에 공유함으로써 혐오표현 탐지에 대한 리서치 과정을 널리 알리는 과정.
pt 준비	5%		아이펠 리서치 과정 마무리를 위한 pt 제작 과정

3-3. 팀장 및 팀원의 역할 분배

순번	주요 담당업무	역할 상세	인원
1	데이터 수집 및 정제	필요한 데이터를 결정하고 수집하는 역할, 기존 데이터셋을 분석하여 부족한 부분을 발견하고 이를 보완할 수 있는 데이터를 제안할 수 있는 사람	최소 2명
2	EDA	데이터의 특성을 이해하고 목표에 맞게 전처리할 수 있는 사람.	1
3	모델개발	더 나은 모델 개발을 위한 수정,개선,보완의 작업을 꼼꼼히 수행하는 자.	1
4	채찍질(🥕)	느슨한 팀플에 적당한 긴장감을 주고 정해진 기한 내에 프로젝트가 완수 될 수 있도록 독려함.	0.3

4. Reference

- [블로그]진행방식에 대한 개괄적인 정리
- [블로그]네이버 ai 클린봇 개발 과정 설명
- [논문]Countering Online Hate Speech: An NLP Perspective
- [논문]Korean Online Hate Speech Dataset for Multilabel Classification - How Can Social Science Improve Dataset on Hate Speech?
- [논문]KoELECTRA를 활용한 챗봇 데이터의 혐오 표현 탐지
- [기사]혐오발전소(국민일보 인터랙티브 기사)
- [블로그]한국어 임베딩 가이드
- [논문][A Survey on Hate Speech Detection using Natural Language Processing](#)
- [웹사이트] [Hate Speech Detection](#)
- [웹사이트] [bert 설명](#)
- [PLM-BERT] [KoBERT\(SKT\)](#)
- [PLM-BERT] [HanBERT](#)
- [PLM-BERT] [KcBERT](#)
- [PLM-BERT] [KR-BERT\(SNU\)](#)
- [PLM-BERT] [Ko-Char-BERT](#)
- [PLM-BERT] [LMKor-BERT](#)
- [PLM-BERT] [DistilKoBERT\(SKT 경량화\)](#)
- [PLM-BERT] [KLUE-BERT](#)
- [PLM-BERT] [KPFBERT](#)
- [PLM-BERT] [KalBert](#)
- [PLM-BERT] [LMKor-Albert](#)
- [PLM-BERT] [KLUE-RoBERTa](#)
- [PLM-BERT] [KoBigBird](#)
- [PLM-ELECTRA] [TUNiB-Electra\(TUNiB\)](#)
- [PLM-ELECTRA] [KoELECTRA](#)
- [PLM-ELECTRA] [KcELECTRA](#)
- [PLM-ELECTRA] [KoCharELECTRA](#)
- [PLM-ELECTRA] [LMKor-ELECTRA](#)
- [PLM-GPT] [LMKor-KoGPT2](#)
- [PLM-GPT] [KoGPT2\(SKT\)](#)
- [PLM-GPT] [KoGPT-Trinity 1.2B\(SKT\)](#)
- [PLM-GPT] [KoGPT\(kakaobrain\)](#)
- [PLM-BART] [KoBART](#)
- [PLM-BART] [Korean-BART](#)