

악플 순화 번역기

목차

1

프로젝트 개요

2

프로젝트 팀 구성 및 역할

3

프로젝트 진행 프로세스

4

프로젝트 결과

5

추후 과제

프로젝트 개요

프로젝트 개요

필요성

'신당역 사건' 피해자에 악성댓글...쏟아지는 2차 가해 어쩌나

피해자 큰아버지 "가슴 아픈 악성 댓글, 드잡이라도 하고 싶다" '인하대 사건' 당시에도 '피해자 행실' 탓하는 2차 가해 비판받아. '신당역 사건' 피해자에 악성댓글...

H 한국경제

슬픔 가시지도 않았는데..."무슨 상관?" 유족 올리는 2차 가해

몇몇 네티즌은 희생자들을 공격하는 글을 자제하자는 유명인들의 SNS로도 단체로 몰려가 악플을 달기도 했다. 한 온라인 커뮤니티에 "이게 추모냐?ㅋㅋㅋ..."

[디지털윤리 부재] 이태원 참사가 드러낸 인터넷 소비 현주소

◇SNS 타고 순식간에 공유된 참사 모습, 희생자 2차 가해 가속화=지난달 29 ... 특히 김 연구소장은 악성 댓글에 대한 국민과 정부, 플랫폼 기업들의...



한겨레

네이버에 뜬 '한겨레' 성범죄 기사 댓글창 닫습니다

성범죄 기사 댓글 창 '2차피해'공간 지적 수용불법촬영 피해자 "댓글 창 방치는 살인 방조" 2차피해 우려 기사 댓글 창 닫기로.

연합뉴스

야후재팬, 악플 방지 강화..."전화번호 등록자만 뉴스 댓글 작성"

(도쿄=연합뉴스) 박상현 특파원 = 일본 주요 포털사이트 야후재팬이 뉴스에 댓글을 작성하려면 휴대전화 번호를 의무적으로 등록해야 하는 제도를 도...

프로젝트 개요


사업성

"포털 성범죄 기사 댓글 2차 가해를 막아주세요" 국민청원, 1만명 ...

May 7, 2021 — 가수 출신 정준영의 불법촬영 피해자가 포털 성범죄 기사의 댓글을 비활성화해달라며 올린 청와대 ... 정준영 피해자 A 씨 "댓글로 심각한 2차 가해..."

네이버·카카오, '건전한 댓글 문화' 조성 팔 걷어붙였다

네이버와 카카오가 댓글 정책을 지속 개편하며 건전한 댓글 문화 확립에 앞장서고 있다.
/더팩트 DB댓글 정책 개편 이후 악성 댓글 자동 감소 추세 나타나[더팩트] ...

 디지털데일리

악성댓글 사회비용 1941억원...“플랫폼 임의임시조치 면책 필요”

[디지털데일리 오병훈 기자]악성댓글 확산을 방지하기 위해서는 플랫폼을 주체로 한 제도·기술적 대응책이 마련돼야 한다는 의견이 제기됐다. 플랫폼 사업자가 적극적...

"악성댓글로 인한 사회적 비용 35조원...제도 개선 필요"

최윤정 방송통신위원회 인터넷이용자정책과장은 디지털 윤리 교육 등 방통위가 현재 운영 중인 다양한 제도를 소개하며 "사업자 자율규제를 강화할 수..."

프로젝트 개요

사업성

악플러는 도태되지 않는다 : 칼럼 : 사설.칼럼 : 뉴스

그렇다고 해서 악플 규제가 아예 의미가 없는 건 아니다. 네이버 연예뉴스에 댓글 기능이 사라진 것만으로도 숨통이 트였다고 말하는 연예인들이 많다...

Feb 7, 2022

악성 댓글, 규제와 차단이 최선인가

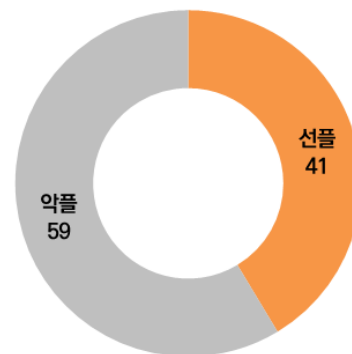
2019년 10월 25일 가수 겸 배우 설리의 극단적 선택을 계기로 악성 댓글에 대한 사회적 공론화 이후 정치권은 '악플방지법'을 발의했다.

Jan 16, 2021

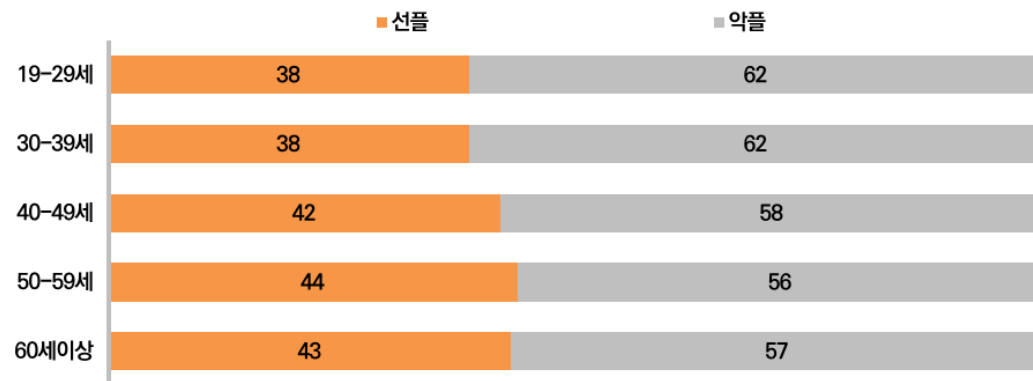
연예뉴스 댓글 사라졌지만 '악성댓글'과의 전쟁은 끝나지 않았다

Aug 6, 2020 — 실제로 경찰청 통계에 따르면 사이버 명예훼손·모욕 발생 건수는 2014년 8880건에서 2019년 1만 6633건으로 약 2배 가량으로 급증했다. 왜곡된 악플 문화 ...

선플 41% VS.악플 59%



2030 악플 비율 응답 62%



질문: 댓글을 선플과 악플로 나누었을 때, 우리 인터넷 웹사이트, SNS, 온라인 동영상 플랫폼의 선플과 악플의 비율은 어느 정도라고 생각하십니까? (선플 %), (악플 %)

비고: base=전체, n=1,000

조사기간: 2019. 11. 15 ~ 18일

한국리서치 정기조사 여론 속의 여론(hrcopinion.co.kr)

프로젝트 개요

구현 내용

무분별한 욕설, 악성 댓글로 인한 2차 가해에 대하여
표현을 해치치 않는 선에서의 필터링과 번역을 하는 시스템을 만들자

프로젝트 개요


활용 라이브러리 및 프레임워크



Colab plus

 PyTorch

PyTorch

 Hugging Face

Hugging Face

 Weights & Biases

wandb



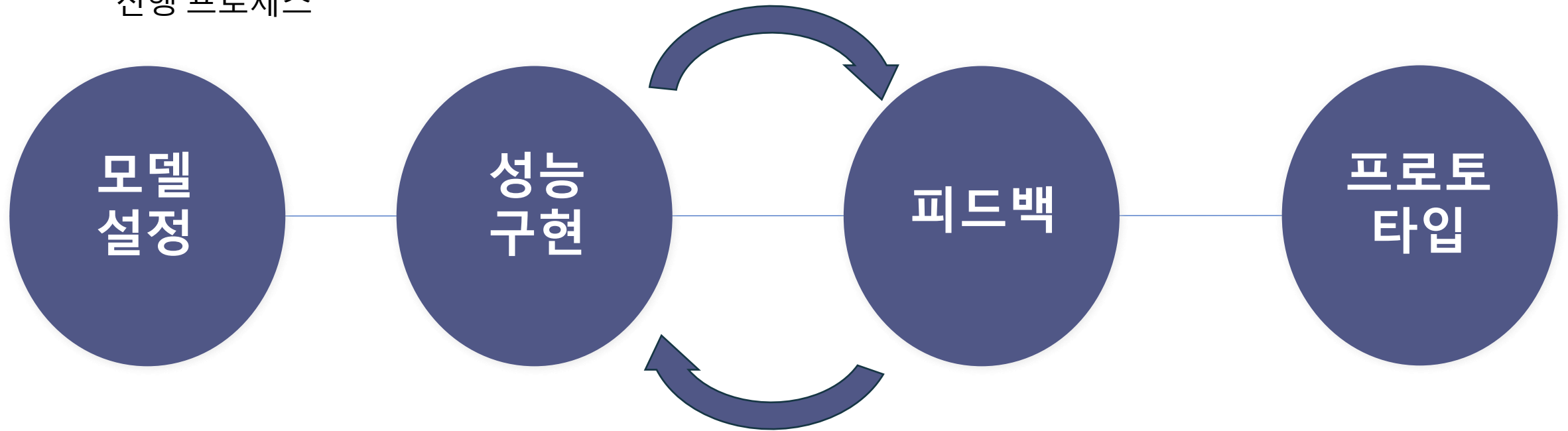
GitHub



streamlit

프로젝트 개요

진행 프로세스



프로젝트 개요

기대 효과

- 악성 댓글에 대한 필터링 자동화
- 표현의 자유를 보장 함과 동시에 악플에 대한 2차 가해를 방지하는 효과
- 악플로 인한 사회적 비용 감소

프로젝트 팀 구성 및 역할

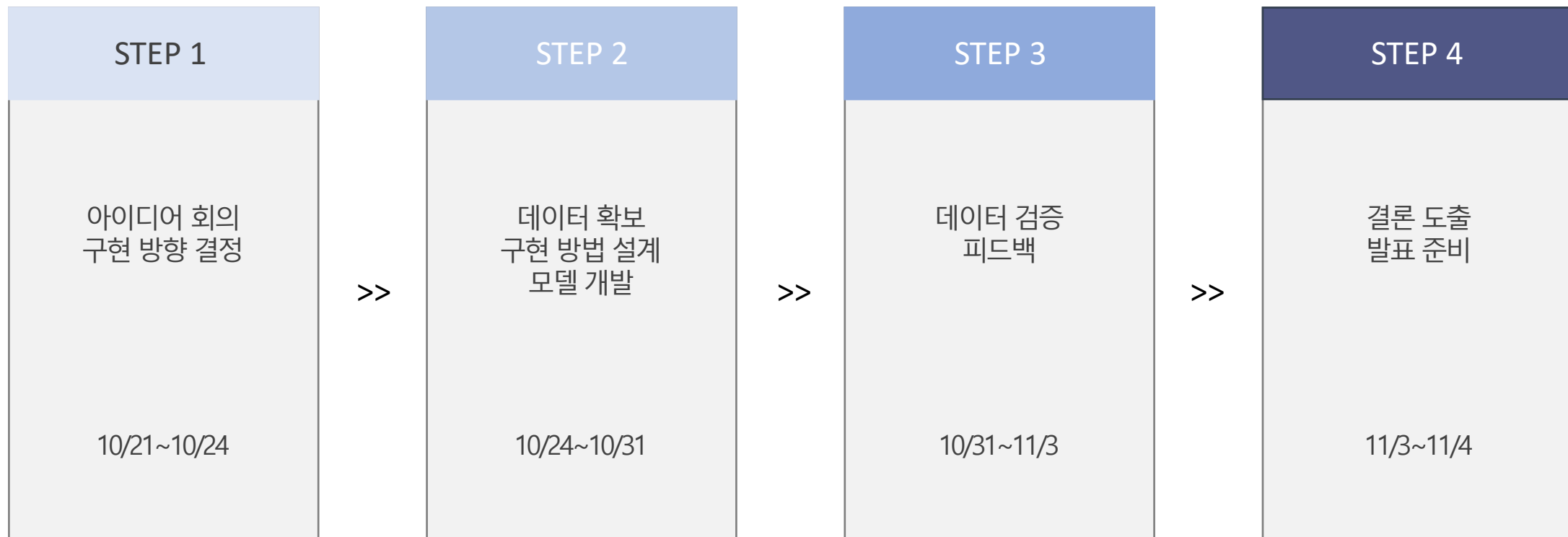
프로젝트 팀 구성 및 역할

이름	역할	
이성미	팀장	관련 논문 분석, 보고서 작성 및 발표
김준형	팀원	관련 논문 분석, data set 탐색
우창	팀원	관련 논문 분석, 프로토타입 (flask), KcELECTRA fine tuning
이지수	팀원	관련 논문 분석, 프로토타입 (streamlit),
최민제	팀원	관련 논문 분석, KcELECTRA fine tuning, Kogpt2 Generater

프로젝트 진행 프로세스

프로젝트 진행 프로세스

일정



프로젝트 결과

프로젝트 결과

학습 데이터 소개

AI hub: 텍스트 윤리 검증 data set

문장 - 453,340문장

비윤리 문장- 251,064문장

-not parallel data set

-비윤리 유형 정보:

"CENSURE": 204,029

"HATE": 69,990

"DISCRIMINATION": 39,885

"SEXUAL": 23,682

"ABUSE": 19,747

"VIOLENCE": 19,562

"CRIME": 8,187

-비윤리 강도의 평균:

1점: 79,137

1점 초과 ~ 2점 미만: 129,230

2점: 26,952

2점 초과 ~ 3점 미만: 10,140

3점: 4,848

2점 미만 비율: 83%

2점 비율: 11%

2점 초과 비율: 6%"

-1점대 데이터 삭제

프로젝트 결과

학습 데이터 소개

Git hub: Korean Hate Speech data set

Labeled - 9,381

-contain_gender_bias, bias, hate labeled:
Hate label만 추가 사용

프로젝트 결과

학습 모델 설명

학습 모델 선정: KcELECTRA

정제되지 않고 구어체 특징에 신조어가 많으며 오탈자 등 공식적인 글 쓰기에 나타나지 않는 표현들이 빈번한 특성을 가진 데이터 셋에 특화
KcELECTRA는 위와 같은 특성의 데이터 셋에 적용하기 위해, 네이버 뉴스 댓글과 대댓글 수집, 토큰나이저와 ELECTRA 모델을 처음부터 학습한 Pretrained ELECTRA모델

선정이유

실험 결과 정제되지 않은 데이터에 관하여 윤리적 비윤리적 댓글을 분류능력이 뛰어남

프로젝트 결과

학습 모델 설명

학습 모델 선정: KoGPT2

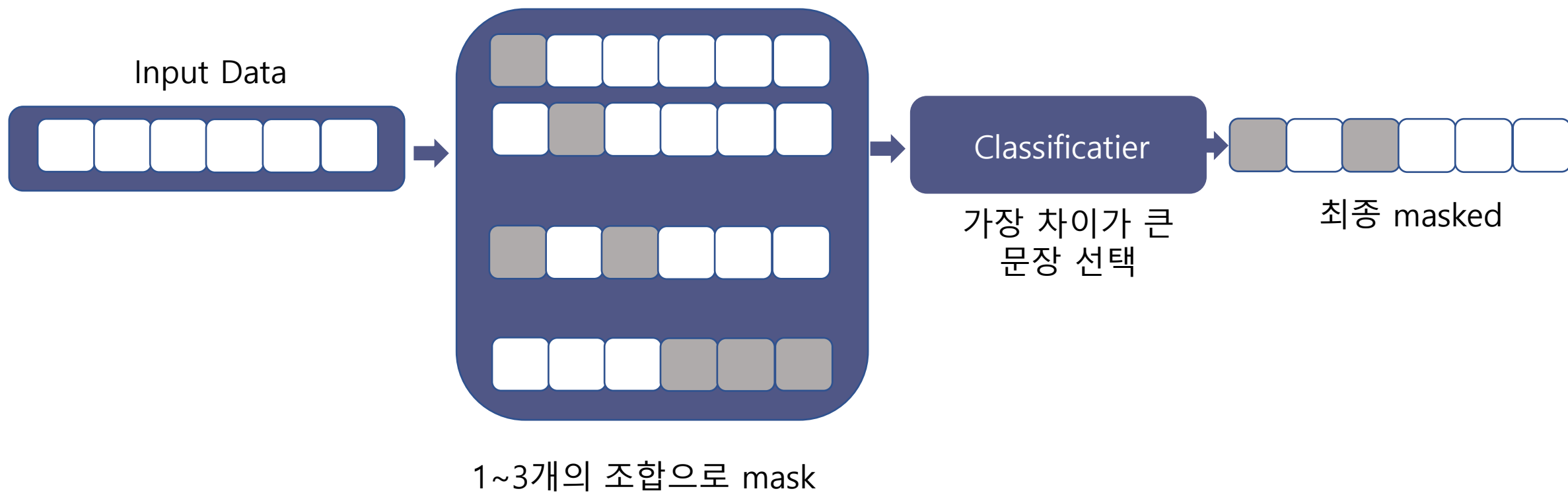
GPT-2는 주어진 텍스트의 다음 단어를 잘 예측 할 수 있도록 학습된 언어모델이며 **문장 생성에 최적화** 되어 있다. KoGPT2는 부족한 한국어 성능을 극복하기 위해 40GB 이상의 텍스트로 학습된 한국어 디코더 언어모델

선정이유

실험 결과 마스킹 된 댓글에 대한 **문장 재구성 능력**이 뛰어남

프로젝트 결과

학습 방식



프로젝트 결과

학습 방식



GPT
model

학습

프로젝트 결과

검증

AI hub test data set classification

accuracy	recall	precision	f1	learning rate	scheduler	batch size	epoch	train_dataset
0.89032396366 46212	0.88288659793 81443	0.87684296747 01614	0.87985440455 57284	5e-5	cosine	256	5	aihub 데이터
0.89166376376 64461	0.88742268041 23712	0.87601767852 98907	0.88168329870 3579	5e-5	constant	256	5	aihub 데이터
0.89166376376 64461	0.86362297496 31812	0.89455699292 165	0.87881785211 16207	3e-5	linear	256	5	korean hate speech 추가
0.89115463972 77527	0.86468335787 92341	0.89265949036 06397	0.87844874019 98922	3e-5	cosine	256	5	korean hate speech 추가
0.89222647980 92124	0.86827687776 14138	0.89192133131 61876	0.87994029850 74627	3e-5	constant	256	5	korean hate speech 추가

프로젝트 결과

프로토타입

- 입력 부분
- 결과 출력
- 비윤리적 댓글 감지 알림

악플 순화기

댓글을 입력해주세요

어디서 이래라저래라 명령질이야 씨발년아 좇까

확인

결과

어디서 이래라저래라 명령질만 하는거야?

비윤리적인 댓글이 감지되었습니다.

악플 순화기

댓글을 입력해주세요

정치글들 제목 보기만 해도 역겨웁

확인

결과

정치글들 보기만 해도 웃기네

비윤리적인 댓글이 감지되었습니다.

프로젝트 결과

라이브 데모

!주의: 번역 전의 예시로 사용된 악플과 욕설이 필터링 없이 나옵니다.

 RUNNING... Stop 

악플 순화기

댓글을 입력해주세요

구름2팀 프로젝트 수고하셨습니다~~

확인



추후 과제

추후 과제

자체 평가 및 보완

일부 가려지지 않는
악플 처리

- 데이터 셋 추가

문맥에 맞지않는
악플 번역 처리

- 번역 이전 댓글과의 유사도 처리

결과 처리 시간 개선

- 조합 기반 masking 방법 개선

상용화 및 배포

- Chrome 확장 프로그램 개발

References

- Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer
- <https://github.com/Beomi/KcELECTRA>
- <https://github.com/SKT-AI/KoGPT2>
- <https://github.com/agaralabs/transformer-drg-style-transfer>
- 마스크 언어 모델 기반 비병렬 한국어 텍스트 스타일 변환

Q&A

감사합니다