# 폭쵕몬 [폭언하는 쵕위는 몬챔아]

NLP 미니프로젝트 최종발표

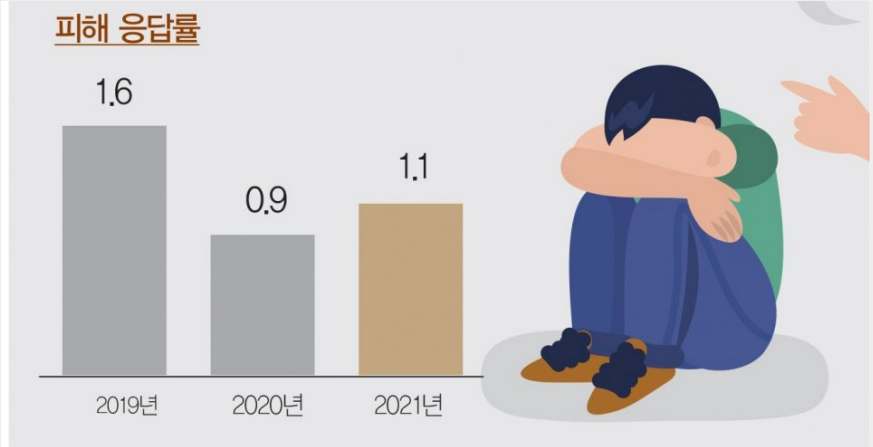
22. 8. 5.

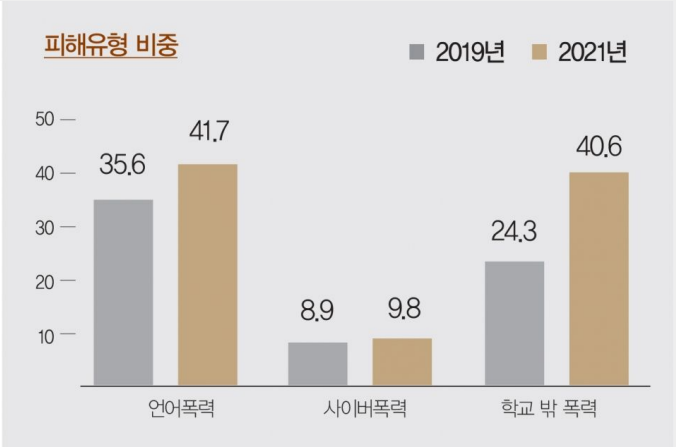
10_조_@10_조_#이정훈#이태훈#정새롬#정준녕

# 목차

# 1. 배경 및 동기 - task 선정 배경 및 이유

**2021년 1차 학교폭력 실태조사 [단위 : %]**




https://news.mt.co.kr/mtview.php?no=2019051316474551634

코로나19 장기화로 원격/비대면 교육 확대 → 사이버 폭력 비중 증가

# 1. 배경 및 동기 - task 선정 배경 및 이유


Google AI

NLP BERT

학생 **화상 상담 서비스** 전면 실시

**온라인 또래상담 관련 플랫폼** 확장 및 콘텐츠 제작·보급

**교원 원격상담 역량 강화** 연수 프로그램 제공

**학교폭력 피해 사실을 알 수 있는 도구의 부재**

유년/청소년기 학생들의 **학교/사이버 폭력 인지를 위한 보조적 시스템** 구현

사이버폭력 등의 폭력적 행태나 장난식으로 협박성 발언하는 등의 것 구별

# 1. 배경 및 동기 - 기대효과

**학교 밖에서 발생하는 학교폭력을 줄일 수 있다**

**학부모 입장에서 사이버 폭력을 확인할 수 있다**
(피해자가 학부모에게 말하지 않는 경우가 많음)

**사이버 폭력에 대한 유무를 정확히 파악할 수 있다**
(어린 아이들의 경우, 피해자 상태 자각 어려움)


**개인적 확인을 통해 개인정보 유출을 방지**

# 2. 실행 계획 - DS관점에서의 문제 정의

- 모델
  - 회귀(Regression)
  - **분류(Classification)**

- NLP 처리
  - **형태소 분석(morphological analysis)**
  - 구문 분석(syntactic analysis)
  - 시멘틱 분석(semantic analysis)
  - 실용 분석(pragmatic analysis)

# 2. 실행 계획 - 데이터셋 소개 및 활용방안

•Github(korean UnSmile Dataset) data(.tsv) [약 **18,000개**, 혐오 10,000, 악플/욕설 4,000 클린 4,600개 데이터] : 사이트

| | 문장 | 여성/가족 | 남성 | 성소수자 | 인종/국적 | 연령 | 지역 | 종교 | 기타 혐오 | 악플/욕설 | clean | 개인지칭 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 일안하는 시간은 쉬고싶어서 그런게 아닐까 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 아동성범죄와 페도버는 기록바 끊어져 영원히 고통 받는다. 무슬림 50퍼 근친이다. ... | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 루나 솔로앨범 나왔을 때부터 머모 기운 있었음 ㅇㅇ Keep o doin 진짜 띵... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

| 항목 | 문장 수 |
|---|---|
| 혐오표현 | 10,139 |
| 악플/욕설 | 3,929 |
| Clean | 4,674 |
| Total | 18,742 |

•AI-Hub(텍스트 윤리검증) data(.txt) [약 **450,000개**, 비윤리 250,000, 윤리 200,000개 데이터] : 사이트

- 2만개 샘플링 후 Unsmile Dataset 합쳐서 학습 계획

{댓글에 빠순이들 몰려와서 즈그 주인님 쉴드치는 꼴 좀 봐이래서 인방충~ 인방충~ 하는거 구나|쉴드가 아니라 국가가 면제해준거야.}

{에로배우들을 데려다가 아이돌을 만드네|쟤네들은 표정이 딱 야동인데..}

{솔직히 여자 48kg이상은 나가 디져야 하는 거 아님|그렇게 따지면 남자 170cm이하도 겨 나오면 안되지|짜져라 이 돼지 꿀꿀아}

•DC 인사이드 크롤링 후 일부 데이터 활용하여 시스템 내 채팅파일로 입력

# 2. 실행 계획 - 데이터셋 소개 및 활용방안

| 데이터셋 이름 | 텍스트 윤리검증 | 한국어 혐오표현 데이터셋 |
|---|---|---|
| 출처 | AI-Hub | Smilegate (Github) |
| 파일형식 | json | .tsv |
| 데이터 수(총합) | 363,154 | 15,005 |
| 윤리 데이터 수 | 162,948 | 3,741 |
| 비윤리 데이터 수 | 200,206 | 11,264 |
| 사용 데이터 수 | 셔플 후 20,000개 샘플링 | 전체 셋 사용 |

→ 총 **35,005 개** dataset 학습 예정, 세부적인 칼럼 통일 필요

**(문장단위 폭력여부 판단 목적)**

# 2. 실행 계획 - 세부 일정

- Project Gantt Chart

| | July | | | | | | | August | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 1 | 2 | 3 | 4 | 5 |
| 프로젝트 기획 | | | | | | | | | | | | |
| 데이터 구축 | | | | | | | | | | | | |
| **Baseline 모델 설계** | | | | | | | | | | | | |
| 프로젝트 중간 발표 | | | | | | | | | | | | |
| 학습 및 성능평가, 모델 수정 | | | | | | | | | | | | |
| **Flask 적용, 웹페이지 구현** | | | | | | | | | | | | |
| 발표자료 작성 | | | | | | | | | | | | |

# 2. 실행 계획 - 프로젝트 프로세스

1. BERT 관련 **선행연구문헌** 탐색/참고하여 **Baseline 모델** 설계
    a. BERT, KR-BERT 관련 키워드 검색
    b. PyTorch Classification 구현 검색

2. 세부 절차
    a. 데이터 전처리
    b. 형태소 분석기 별 소요시간/성능 평가
    c. 하이퍼파라미터 별 성능 영향여부 확인
    d. 학습데이터 증가에 따른 성능변화 경향 파악

3. 웹페이지 구현

# 3. 진행 결과

- 데이터 전처리
  - 윤리 데이터셋 : json → csv로 **파일 형태 변환**
    - 각 데이터 **컬럼 명**을 clean, violence로 **통일**
    - 윤리 데이터 셋에서 **clean 과 violence 비율에 맞춰서** 데이터를 통합
    - 이후 랜덤으로 **train, valid 데이터 나누어서** 저장
    - clean, violence의 값이 **명확하지 않은 문장 재처리**
      - 항상 같은 값을 가지도록 random.seed의 값 통일
  - 파일 변환, **순수 한글 데이터 추출**을 위해 특수문자, 이모지, 영문 제거

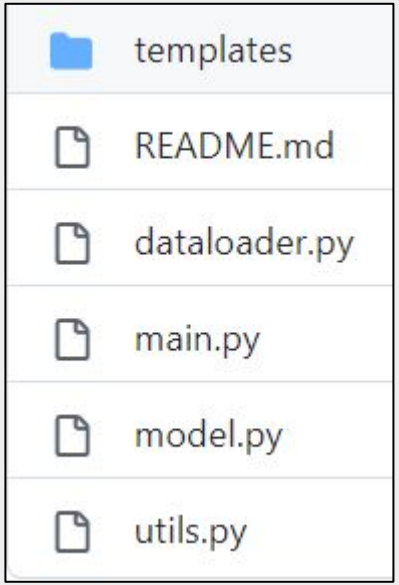
| 데이터 가공 | 폭력 (1) | 비폭력 (0) |
|---|---|---|
| 윤리 세트 약 20,000개 | T | F |
| smilegate 약 18,000개 | clean 외 | clean |

**데이터 통일 기준**

# 3. 진행 결과

- Baseline 모델(bert-base-uncased) 설계
  - Baseline 참고
    - https://arxiv.org/abs/1810.04805
    - https://luv-bansal.medium.com/fine-tuning-bert-for-text-classification-in-pytorch-503d97342db2
    - https://mccormickml.com/2019/07/22/BERT-fine-tuning/
  - KR-BERT 참고
    - snunlp/KR-BERT: KoRean based BERT pre-trained models (KR-BERT) for Tensorflow and PyTorch (github.com)

➪ 파이썬 **스크립트로 모듈화** 진행




모델 트리 구조

# 3. 진행 결과

- 형태소 분석기 별 소요시간/성능 평가

- 목적 : 최적의 형태소 분석기 찾기 [ 0 : clean, 1 : violence ]
  - 1차 : **Konlpy 간 성능 비교** (okt, kkma, komoran, hannanum, mecab)
  - 2차 : mecab 내 사용 **품사간** 성능 비교(체언, 용언, 관형사, 조사, 부사)
  - 3차 : **mecab, khaiii(카이), kiwi** 간 성능 비교
  - 4차 : **BERT vs KR-BERT** 모델 성능 비교
  - 번외 : **형태소 분석기** 에러 원인 분석

⇨ 성능 비교 후 **최고 모델을 확인** 후 **웹페이지 구현 시 반영**

# 3. 진행 결과

- Pretrained-Model, max length,
  batch size 별 성능 영향여부 확인

  - BERT : bert-based-uncased [1]
  - KR-BERT : snunlp/KR-BERT-char16424 [2]

    - 작은 vocabulary, 적은 parameters,
      적은 training dataset으로도 좋은 성능을 보임
    - 한국어의 특성을 반영하여,
      **어근과 접사를 나누어** 학습시킴

| | Mulitlingual BERT | KR-BERT character BidirectionalWordPiece | KR-BERT sub-character WordPiece |
|---|---|---|---|
| 냉장고<br>nayngcangko<br>"refrigerator" | 냉#장#고<br>nayng#cang#ko | 냉장고<br>nayngcangko | 냉장고<br>nayngcangko |
| 춥다<br>chwupta<br>"cold" | [UNK] | 춥#다<br>chwup#ta | 추#ㅂ다<br>chwu#pta |
| 뱃사람<br>paytsalam<br>"seaman" | [UNK] | 뱃#사람<br>payt#salam | 배#ㅅ#사람<br>pay#t#salam |
| 마이크<br>maikhu<br>"microphone" | 마#이#크<br>ma#i#khu | 마이크<br>maikhu | 마이크<br>maikhu |

| Model | Masked LM Accuracy |
|---|---|
| KoBERT | 0.750 |
| KR-BERT character BidirectionalWordPiece | **0.779** |
| KR-BERT sub-character BidirectionalWordPiece | 0.769 |

# 형태소 분석기 실험 및 과정(1차)

- Konlpy간 성능 비교 (okt, kkma, komoran, hannanum, mecab)

| | f1-score | | | | 속도비교 (1 epoch) |
|---|---|---|---|---|---|
| | 1 epoch | | 4 epoch | | |
| | 0 | 1 | 0 | 1 | |
| **원본문장** | 0.48 | 0.81 | - | - | |
| **okt** | 0.46 | 0.81 | 0.57 | 0.81 | 1분 15초 |
| **kkma** | 0.47 | 0.80 | 0.60 | 0.81 | 40초 |
| **komoran** | 0.54 | 0.79 | 0.61 | 0.81 | 14분 30초 |
| **hannanum** | 0.47 | 0.80 | 0.59 | 0.81 | 2분 30초 |
| **mecab** | 0.41 | 0.81 | 0.62 | 0.81 | 5초 |

➢ 1 → 4 에폭 증가 시, 0 에 대한 **성능이** 개별적으로 약 **7 ~ 20 % 까지 증가**

➢ **mecab 형태소** 분석기가 가장 **빠른 속도**로 진행

# 형태소 분석기 실험 및 과정(2차)

- mecab 내 사용 품사간 성능 비교(체언,용언,관형사,조사)

| mecab | f1-score | |
|---|---|---|
| | 1 epoch | |
| | 0 | 1 |
| 체언, 용언 | 0.41 | 0.81 |
| 체언, 용언, 관형사, 부사 | 0.47 | 0.80 |
| 체언, 용언, 관형사, 부사, 조사 | 0.47 | 0.81 |

➢ **품사**가 많을수록 **0에서** 학습성능 향상

# 형태소 분석기 실험 및 과정(3차)

- mecab vs khaiii(카이) vs kiwi 간 성능 비교

| | 4 epoch | | | 속도비교 |
|---|---|---|---|---|
| | f1-score | | accuracy | |
| | 0 | 1 | | |
| **mecab** | 0.62 | 0.81 | 0.75 | 5초 |
| **khaiii** | 0.6 | 0.81 | 0.74 | 24초 |
| **kiwi*** | 0.62 | 0.8 | 0.74 | 44초 |

➢ 성능 부분에서는 큰 차이가 없음을 알 수 있다.

➢ mecab 형태소 분석기가 가장 빠른 속도로 진행

*** Kiwi : 지능형 한국어 형태소 분석기(Korean Intelligent Word Identifier) : 빠른 속도와 범용적인 성능의 형태소 분석기**

# 형태소 분석기 실험 및 과정(4차)

- BERT vs KR-BERT 모델 성능 비교

| 형태소 분석기 | BERT | | | | | | KR-BERT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 epoch | | | 4 epoch | | | 1 epoch | | | 4 epoch | | |
| | f1-score | | accuracy | f1-score | | accuracy | f1-score | | accuracy | f1-score | | accuracy |
| | 0 | 1 | | 0 | 1 | | 0 | 1 | | 0 | 1 | |
| **mecab** | 0.47 | 0.81 | 0.72 | 0.62 | 0.81 | 0.75 | 0.68 | 0.85 | 0.79 | 0.69 | 0.84 | 0.79 |
| **khaiii** | | | | 0.6 | 0.81 | 0.74 | 0.71 | 0.86 | 0.82 | 0.71 | 0.85 | 0.8 |
| **kiwi** | | | | 0.62 | 0.8 | 0.74 | 0.73 | 0.86 | 0.82 | 0.71 | 0.86 | 0.81 |

- ➢ 4 epoch 기준, BERT보다 **KR-BERT의 성능이 우수**
- ➢ KR-BERT 모델 : mecab보다 **kiwi의 성능이 우수**

# 형태소 분석기 실험 및 과정(번외)

- 형태소 분석기 에러 발생

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 2278 |
| 1 | 0.67 | 1.00 | 0.80 | 4586 |
| accuracy | | | 0.67 | 6864 |
| macro avg | 0.33 | 0.50 | 0.40 | 6864 |
| weighted avg | 0.45 | 0.67 | 0.54 | 6864 |

문제 발생 예시

➢ 모델이 **모든 예측을 1로** 하는 문제 발생

➢ okt, kkma, komoran, hannanum 형태소 분석기에서 모두 문제 발생

# 형태소 분석기 실험 및 과정(번외)

- 원인 찾기

| | 원인 찾기 | 결과 |
|---|---|---|
| 1 | 컬럼 타입이 다른가 | no |
| 2 | 파일 저장하는 과정에서 문제 발생? | no |
| 3 | 텍스트 전처리 + 형태소 분석기 사용 | 문제 발생 |
| 4 | 텍스트 전처리만 사용 | 문제 발생 |
| 5 | 형태소 분석기만 사용 | no |
| 6 | 에폭이 낮아서(학습이 부족해서) 나타날까?<br>(에폭을 1에서 4로 늘려도 문제는 동일) | no |

➢ **텍스트 전처리 과정에서** 문제가 발생함을 파악

➢ 텍스트 **전처리 코드를 수정**하여 문제를 해결

# Pretrained Model, max length, batch size 별 성능 영향여부 확인

| | f1-score | | | |
|---|---|---|---|---|
| | BERT | | KR-BERT | |
| max length/ batch size | 0 | 1 | 0 | 1 |
| 128 / 32 | 0.41 | 0.81 | 0.74 | 0.87 |
| 128 / 16 | 0.44 | 0.81 | 0.73 | 0.86 |
| 256 / 32 | 0.41 | 0.81 | 0.74 | 0.87 |
| 256 / 16 | 0.36 | 0.81 | 0.73 | 0.86 |

- KR-BERT 모델이 학습 성능이 전반적으로 우수, **0에서 뛰어난 향상** 보임 (30% ↑)
- max length : 성능차이 미미, 그에 반해 **학습시간이 128에서 2배 빠름**
- batch size : 전반적으로 **약간의 성능 향상** (BERT : -3 ~ 5%, KR-BERT : 1%)

# 학습데이터 증가에 따른 성능변화 경향 파악

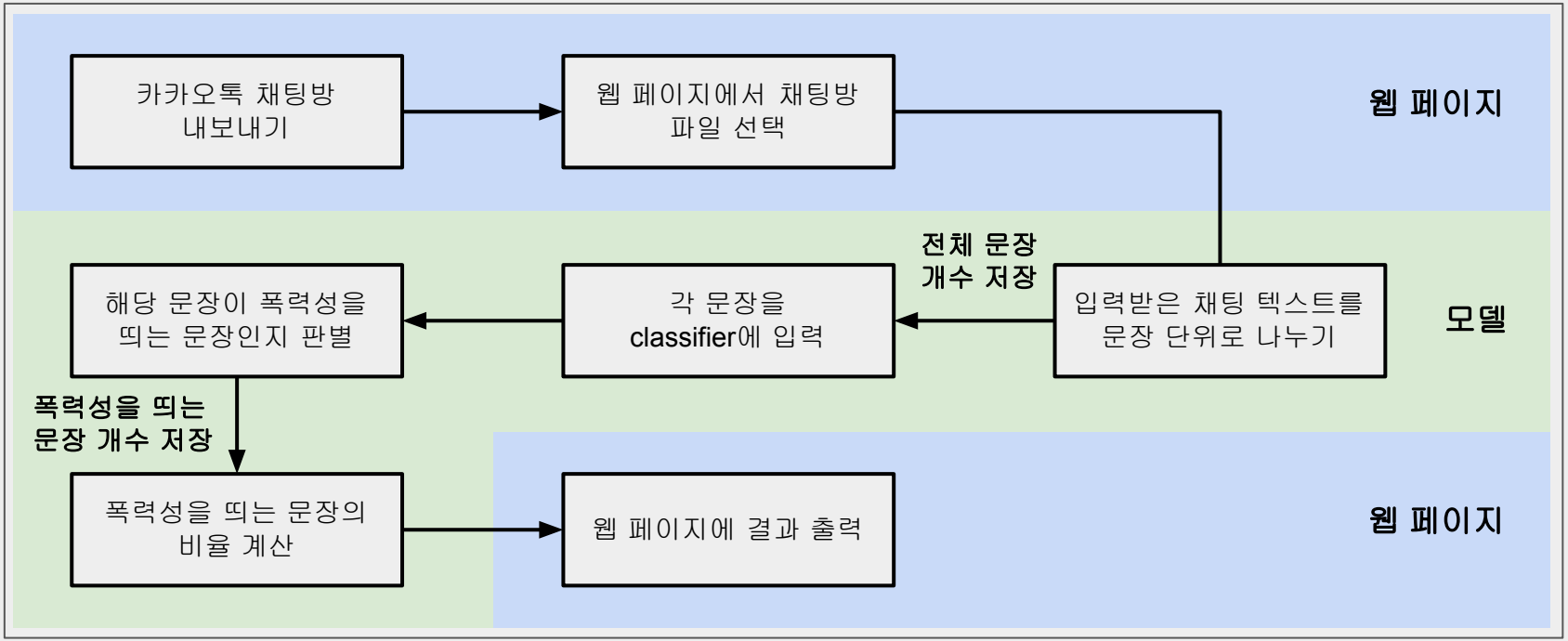
| | BERT, 4 epoch, batch size 32, max length 128 | | | | | |
|---|---|---|---|---|---|---|
| | baseline (10%) | | Data 50% | | Data 100% | |
| **Target** | **0** | **1** | **0** | **1** | **0** | **1** |
| **f1-score** | 0.41 | 0.81 | 0.72 | 0.78 | 0.74 | 0.79 |
| **소요시간** | 약 32분 | | 약 2시간 51분 | | 약 5시간 54분 | |
| **환경** | Colab (Tesla T4) | | Colab Pro (Tesla T4) | | | |

➢ 데이터 셋 증가에 따라, 성능 향상하는 경향 확인, 소요시간도 비례해서 증가

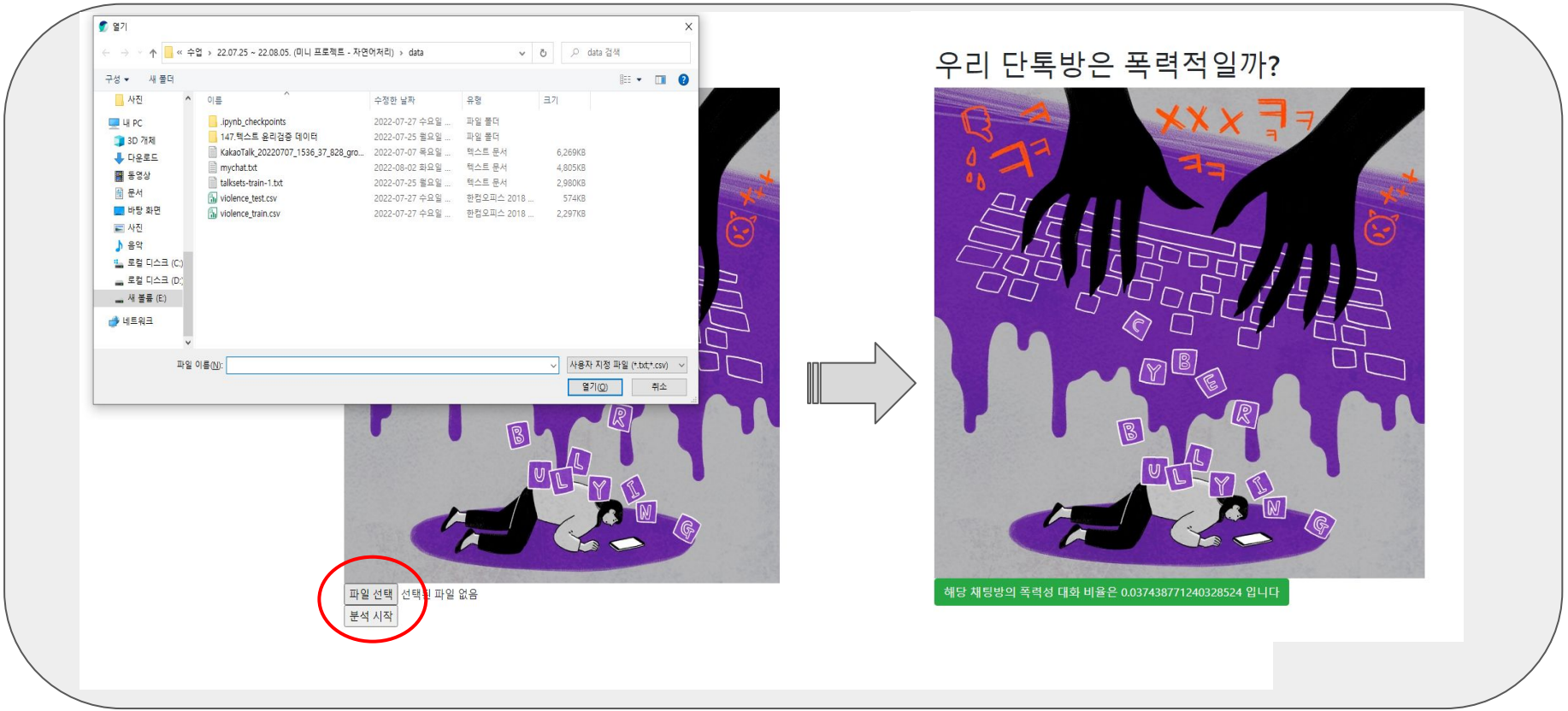
⇨ **Data 100%** 모델 채택

# 3. 웹페이지 구현 순서도

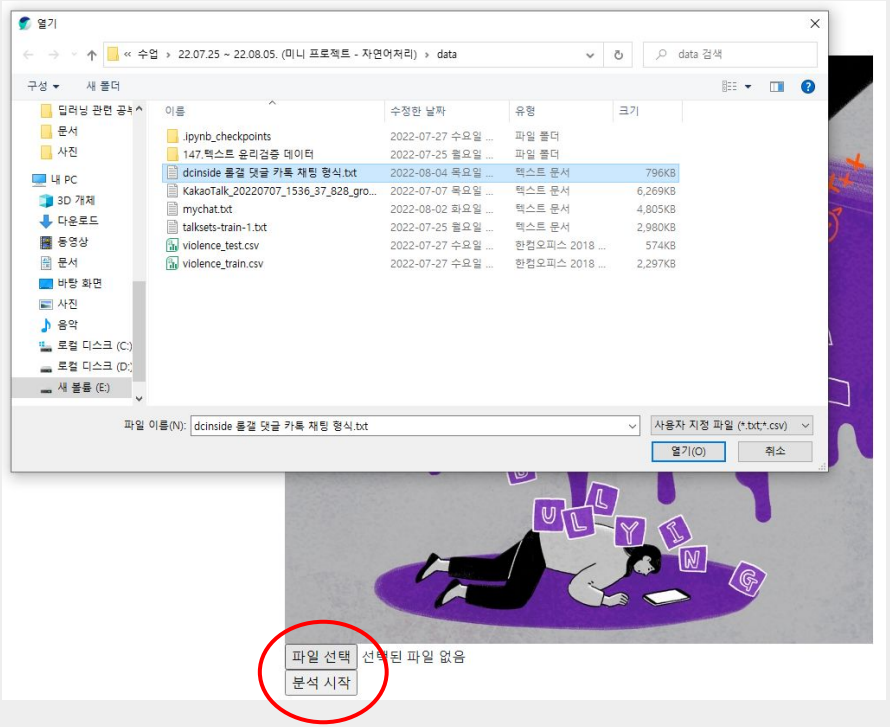
- Flask 구축 (기본 Back/Frontend)

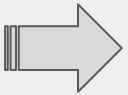


# 예시 화면


열기
« 수업 › 22.07.25 ~ 22.08.05. (미니 프로젝트 - 자연어처리) › data
data 검색
구성
새 폴더
사진
내 PC
3D 개체
다운로드
동영상
문서
바탕 화면
사진
음악
로컬 디스크 (C:)
로컬 디스크 (D:)
새 볼륨 (E:)
네트워크
이름
수정한 날짜
유형
크기
.ipynb_checkpoints 2022-07-27 수요일 ... 파일 폴더
147.텍스트 윤리검증 데이터 2022-07-25 월요일 ... 파일 폴더
KakaoTalk_20220707_1536_37_828_gro... 2022-07-07 목요일 ... 텍스트 문서 6,269KB
mychat.txt 2022-08-02 화요일 ... 텍스트 문서 4,805KB
talksets-train-1.txt 2022-07-25 월요일 ... 텍스트 문서 2,980KB
violence_test.csv 2022-07-27 수요일 ... 한컴오피스 2018 ... 574KB
violence_train.csv 2022-07-27 수요일 ... 한컴오피스 2018 ... 2,297KB
파일 이름(N):
사용자 지정 파일 (*.txt;*.csv)
열기(O)
취소
파일 선택
선택된 파일 없음
분석 시작
우리 단톡방은 폭력적일까?
해당 채팅방의 폭력성 대화 비율은 0.037438771240328524 입니다

# 예시 화면


열기
« 수업 › 22.07.25 ~ 22.08.05. (미니 프로젝트 - 자연어처리) › data
data 검색
구성
새 폴더
이름
수정한 날짜
유형
크기
.ipynb_checkpoints
147.텍스트 윤리검증 데이터
dcinside 롤갤 댓글 카톡 채팅 형식.txt
KakaoTalk_20220707_1536_37_828_gro...
mychat.txt
talksets-train-1.txt
violence_test.csv
violence_train.csv
파일 이름(N): dcinside 롤갤 댓글 카톡 채팅 형식.txt
사용자 지정 파일 (*.txt;*.csv)
열기(O)
취소
파일 선택
선택된 파일 없음
분석 시작




우리 단톡방은 폭력적일까?
해당 채팅방의 폭력성 대화 비율은 80.71% 입니다

# 웹 페이지 개선할 사항

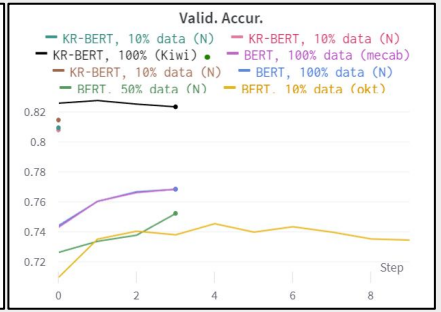
➢ 채팅파일 선택과 결과 출력을 동일 페이지 내 구현하도록 개선 필요
- 현재, **채팅파일 선택**과 **결과 출력** 페이지가 별도 존재,
  결과 확인 후 파일 선택을 위해 채팅파일 선택 페이지로 이동

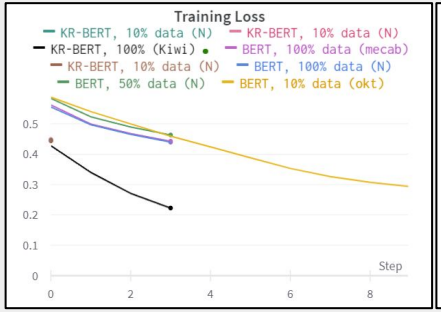
➢ 결과 산정 소요시간 단축
- 채팅파일 약 6MB(문장 12.2만건) 기준 **40분 소요**
- 반복문으로 **각 문장을 하나씩 처리**하기 때문에 소요시간이 증가

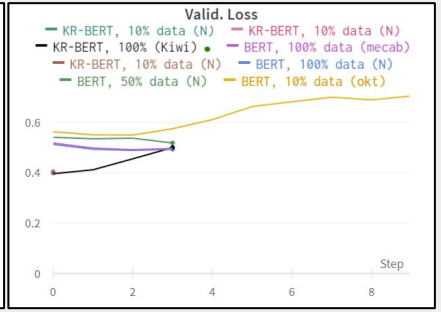
∴ 모델입력 데이터 형식 : **문장 ➪ 채팅파일 단위** 입력 및 비율 계산으로 개선

➢ 페이지 디자인 개선 및 웹 서비스 적용 (AWS)

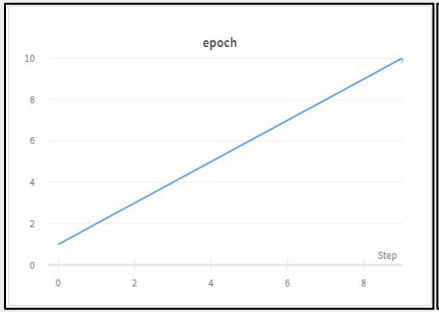
# WandB 연동, 모델 별 성능 비교

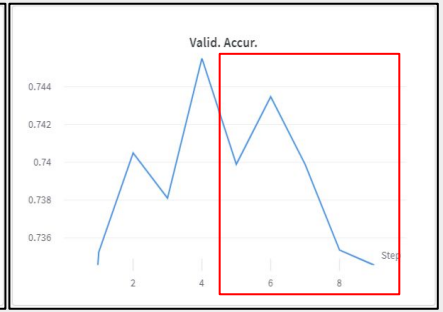



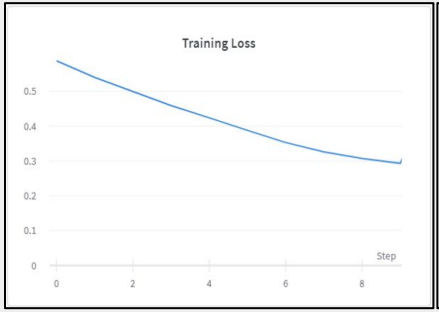



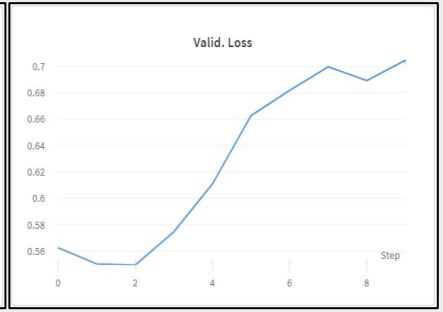




epoch 수, 데이터 셋, 형태소 분석기 처리 등 다중모델 간 비교




overfitting 되는 경향 확인 (valid accuracy 감소, **10 epochs**)

➢ WandB 활용, 개별모델 학습 경과 및 경향 확인

# 요약 및 향후 계획

- 요약
  - 사이버 폭력의 증가에 따라, 이를 **예방하기 위한 실질적인 장치 필요**
  - **BERT모델 기반**으로 채팅내역 내 **폭력성인지 시스템** 구현
  - 형태소 분석기 : BERT : mecab / KR-BERT : kiwi
    - 성능도 준수하며, 속도 측에서 큰 장점을 가짐
  - KR-BERT 사용 시, **어근/접사의 분리 학습으로, BERT대비 성능 향상**
  - 선정된 우수 조건 반영 및 학습 시 최종적으로 **82 %** 성능 확보

- 향후 계획
  - 데이터 셋 규모/성능의 비례관계 확인, 향후 성능개선을 위해 **관련 데이터를 수집** 계획 중
  - WandB sweep 활용, **하이퍼 파라미터(learning rate, eps) 탐색** 진행
  - early stopping, save best weights 구현 및 PyTorch Ignite 적용
  - KR-BERT 모델 저장 및 불러오기 구현

# Questions ?

Thanks !

Example codes will be available at

https://github.com/NLP-yd10/CheckViolence.

End