저한테 대체 뭘 바라시는 겁니까

2025.5.7

박민경 , 이주연, 김민지, 김지환

목 차

프로젝트 목표

문제정의

train.csv

협박	갈취	직장 내 괴롭힘	기타 괴롭힘
폭력적이거나 위협적인 언어를 통해 상대에게 두려움을 주는 대화	강압적으로 금전이나 물건을 요구하는 대화	업무 외적인 요구나 상사의 횡포, 인격모독 등이 포함된 직장 내 부당한 요구나 언행	성적 발언, 외모 비하, 따돌림 등 기타 다양한 형태의 괴롭힘
지금 너 스스로를 죽여 달라고 애원하는 것인가	어이 거기 예 너 말이야 너 이리 오라고	나 이틀 뒤에 가나다 음식점 예약 좀 해줘. 가나다 음식점이요?	너 되게 귀여운 거 알지? 나보다 작은 남자는 첨 봤어.

test.csv

일반 대화
괴롭힘과 무관한 일상적인 대화
오늘 날씨 어때? 날씨 맑고 기온도 적당해서 좋아. 그러게, 나도 바람 쐬고 싶어.

텍스트로 주어진 대화를 다섯 가지 클래스 중 일반 대화를 정확하게 분류하는 딥러닝 모델을 개발

문제 접근

DATA

- 데이터 내용 파악
- 데이터 EDA
- 데이터 생성
- 데이터 Augmentation

MODEL

- 토크나이저 비교
- 모델 비교
- 실험
- 모델 선정

데이터 전처리

EDA

데이터의 형태

각 발화마다 줄 바꿈(\n)으로 구분
공백 라인 존재

- 상황마다 다른 의미를 가진다 : 침묵한 경우 이거나, 결측 데이터인 경우로 판단

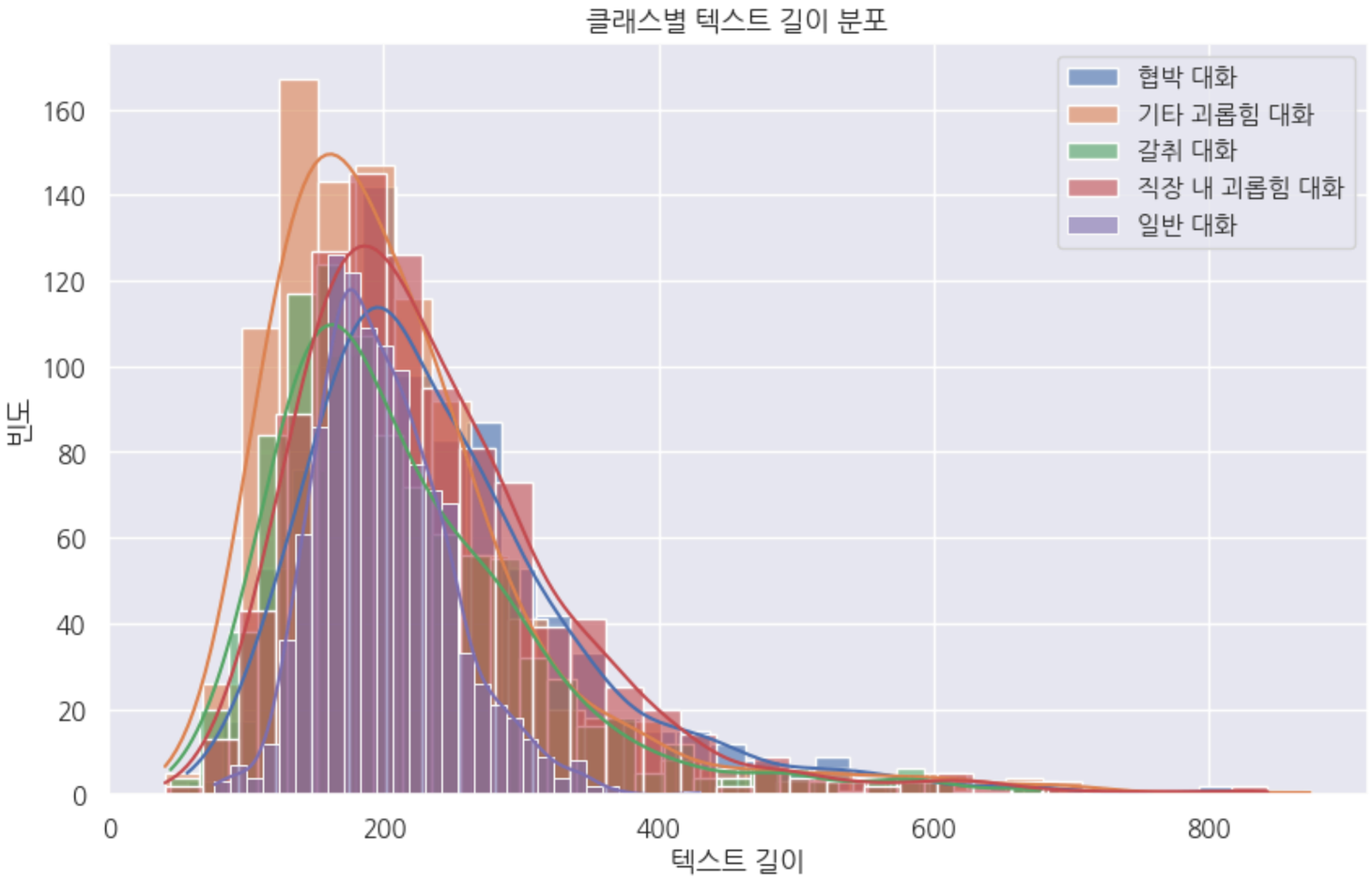
재 저거 다 고치고 들어온 거라며?
진짜? 어쩐지 저렇게 예쁠 수가 없지
무슨 소리야?
아니 너 언제는 얼굴 하나도 안고친거라며.
아니야.안고쳤어
뭐라는거야 내가 니 중학고 고등학교 졸업 사진을 봤는데
언제 그렇게 고쳤냐? 진짜 용됐다 아주?
아닌데 그냥 살빠져서 그런거야.
놀고 있네 이게 살만 빠진다고 나올 얼굴이냐!!! 어디서 개사기를 쳐!!!
니가 사진 뿌리고 다닌거야?
어머! 무슨소리야 설마 이걸 나만 알고 있다고 생각한건 아니지?
우리 아닌데

데이터의 통계량

클래스 별 텍스트 길이 통계량 대체로 비슷함

클래스	최대 길이	최소 길이	평균 길이
협박 대화	818	57	246.07
기타 괴롭힘 대화	874	41	210.08
갈취 대화	678	45	216.19
직장 내 괴롭힘 대화	843	41	237.56
일반 대화	430	77	200.18

EDA 시각화



워드클라우드

감탄사(아, ㅇㅇ ㅑ 아니, ...)등의 비중이 매우 높음

→ 불용어 처리가 일부 필요해 보인다.

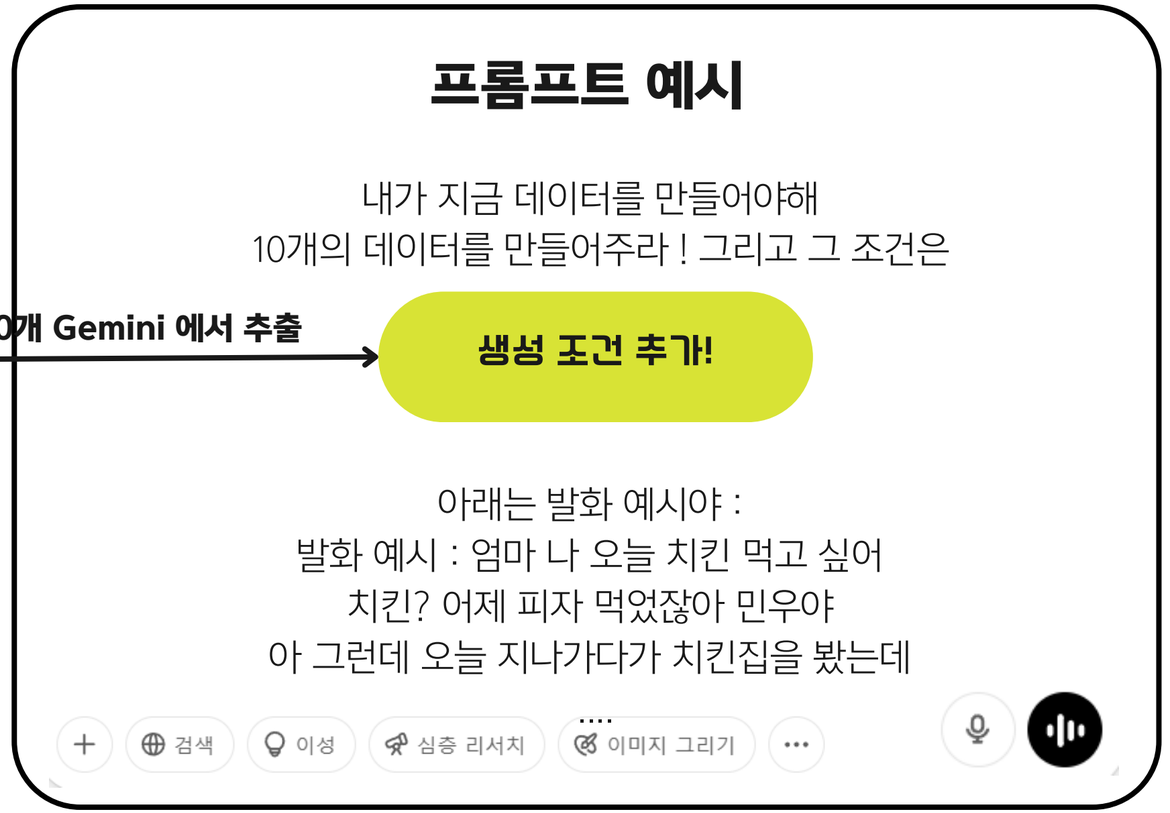
데이터 생성

Generative Model을 사용한 데이터 생성 진행

사용 Model : Gemini Claude ChatGPT

자연스러운 대화: 서로 질문하고 답변하는 느낌! (무작위 문장 나열 X)
말투: 반말/존댓말 섞임 (상황에 맞게)
주제: 20개)
대화 인원: 2명~3명 (A, B 또는 A, B, C)
턴 수: 평균 10턴 (최소 8~최대 12턴 정도)
발화 길이:
최대 글자 수: 200자
평균 글자 수: 20자
최대 단어 수: 60단어
평균 단어 수: 6단어
최소 글자 수: 0자 (가끔 공백 발화 허용)
형식:
CSV 첫 줄: idx,class,conversation
모든 idx: 4부터 시작 (또는 4로 고정)
모든 class: "일반 대화"
대화 내용은 줄바꿈(\n)으로 구분
공백발화는 줄바꿈(\n)으로 표현
발화 하나 생성할 때마다 이전 발화를 보고 문맥을 이어서 생성

대화 주제 1000개 Gemini 에서 추출



총 생성 데이터 (일반대화) : 1123개

Augmentation

2가지 방식의 Augmentation 실험

1. KorEDA
2. GPT-4.1-mini 를 활용한 generative augmentation

모든 데이터 증강 방식은 dataleakage를 피하기 위해 split후 사용, train, validation set에 대해서 증강함.
두 방식 모두 데이터 한 개당 두개 증강 생성

Augmentation

1. KorEDA
- SR: Synonym Replacement, 특정 단어를 유의어로 교체
- RI: Random Insertion, 임의의 단어를 삽입
- RS: Random Swap, 문장 내 임의의 두 단어의 위치를 바꿈
- RD: Random Deletion: 임의의 단어를 삭제

카이스트에서 만든 Korean WordNet을 사용한 EDA 방식 사용.

선정 이유 :

1. 한글용으로 최적화됨
2. SR, RI, RS, RD 등으로 간단하고 빠르게 증강 가능
3. 무작위 규칙 기반 대체/삽입/삭제 해 모델이 문장 구조 다양성에 익숙해지도록 한다

Augmentation

```
base_prompt = """
다음의 주어진 문장들의 의미와 어조는 동일하되, 2개의 다른 표현으로 변형해주세요.
결과는 문자열 리스트로 출력해주세요.

아래의 예시는 참고용입니다.
입력:
요즘 뭐 하고 지내?\n그냥 과제랑 알바하면서 바쁘게 지내고 있어. 너는?\n나도 비슷하지. 근데 혹시 이번 주 토요일에 시간 돼? 영화나 보러 갈래?\n좋아! 어제 개봉한 영화 보러갈까?\n영화 인터스텔라 말하는 거지? 난 좋아.\n좋아 그럼 토요일 낮에 보자. 12시 어때\n그래 그때 영화관에서 보자.

출력:
[
"요즘 어떻게 지내?\n그냥 과제랑 알바 병행하느라 정신없이 지내고 있어. 너는?\n나도 뭐, 별반 다르지 않아. 그런데 이번 주 토요일에 시간 괜찮아? 영화 하나 볼까 해서.\n좋지! 어제 개봉한 영화 보러 가는 건 어때?\n혹시 인터스텔라 말하는 거야? 나 완전 좋아.\n그럼 토요일 낮에 보자. 12시쯤 어때?\n좋아, 그때 영화관 앞에서 만나자.",
"최근에 뭐하고 지냈어?\n과제랑 알바하느라 하루하루 정신없이 보내고 있어. 너는 어때?\n나도 마찬가지야. 근데 이번 주 토요일에 영화나 같이 볼래?\n좋아! 어제 개봉한 거 보자.\n인터스텔라 말하는 거지? 완전 좋아!\n그럼 토요일 낮 12시에 만나자.\n응, 영화관 앞에서 보자!"
]

이 예시를 바탕으로 다음의 주어진 문장들의 표현을 다르게 변형해주세요.
입력:
"""
```

2. GPT-4.1-mini 를 사용한 데이터 증강
   - Train, Validation set의 각 데이터에 대해서 2개씩 증강 요청
   - 1 shot prompting을 통한 생성 예시 제시
   - train 3477 → 10373 개 데이터로 증강(중복 처리 후)
   - val 745 → 2256개 데이터로 증강(중복 처리 후)

선정 이유 :

1. 프롬프트 기반 문장 생성 가능
2. 의미 유지 기반의 문장 생성 : 감정 표현이 유지된 상태로 문장 다변화 가능

Augmentation

아빠 나 급하게 살 거 생겼는데 빨리 100만원만 보내줘
어떤 걸 사는데 100만원이나 필요하니
그건 알 바 없고 빨리 보내주기나 해
아빠가 요즘 상황이 안좋은 거 알잖아
아 거짓말 치지마 아빠 돈 벌잖아 그 정도는 있을 거 아니야
아빠도 주고 싶은데 정말 상황이 여의치 않아서 그래
아 진짜 짜증나게 헛소리하지 말고 빨리 보내줘
아빠가 미안해 못 보내줘
그래 그럼 아빠 서랍에 있는 통장 가져간다
아들 아빠가 얼른 구해줄게 조금만 기다려주겠니

아빠, 나 급하게 돈이 필요한데 100만원만 빨리 보내줄 수 있어?
100만원씩이나 어디에 쓰는데?
그건 중요하지 않아, 그냥 빨리 보내줘.
아빠 요즘 힘든 거 알잖아.
아니야, 거짓말 하지 마. 아빠 돈 벌잖아, 그정도는 있을 거야.
아빠도 주고 싶은데 진짜 상황이 안 좋아서 그래.
아, 정말 화나게 하지 말고 빨리 보내달라고.
미안해, 지금은 못 보내줘.
그래, 그럼 아빠 서랍에 있는 통장 가져갈게.
아들, 내가 최대한 빨리 구해줄게, 조금만 기다려줘.

아버지, 급하게 100만원이 필요해요. 빨리 보내줄 수 있어요?
100만원이나 필요한 이유가 뭐야?
그건 신경 쓰지 말고 빨리 보내줘요.
아빠 요즘 상황 어렵다는 거 알면서.
아니야, 거짓말하지 마. 아빠 돈 버는 거 알잖아, 그 정도는 있잖아.
아빠도 주고 싶은 마음은 있지만 상황이 여의치 않아.
진짜 귀찮게 하지 말고 얼른 보내줘.
미안해, 지금은 못 보내겠어.
알겠어, 그럼 아빠 서랍에서 통장 가져갈게.
아들, 금방 돈 구해볼게, 조금만 기다려줘.

모델

모델 서칭

모델 서칭을 위한 데이터 특징

- 한국어 텍스트 분류 작업에 적합
- 감정 분석, 의도 분류, 혐오 표현 탐지 등과 유사한 특성
- 한국어에 특화된 사전학습 모델을 사용하는 것이 효과적

실험 진행한 모델

klue/bert-base	형태소 기반의 토크나이징 방법을 사용하여 한국어의 특성을 반영
monologg/koelectra-base-v3-discriminator	아키텍처 기반의 Discriminator로, 입력 문장에서 일부 토큰을 가짜로 교체하여 이를 판별하는 방식으로 학습
KoElectra-base-v3-discriminator	ELECTRA는 'Replaced Token Detection' 방식을 사용하여, 생성된 가짜 토큰이 실제인지 판별하는 방식으로 학습
klue/RoBERTa	기존 RoBERTa 모델은 영어 중심의 데이터셋으로 학습되었으나, KLUE/RoBERTa는 한국어에 최적화된 사전 학습을 통해 한국어 이해 능력을 향상
KoBERT (SKT)	한국어 위키백과와 뉴스 데이터를 기반으로 학습

주요 시도

모델과 토크나이저 적용 실험

데이터에 따라 결과가 달라질 수 있으므로 여러 모델과 토크나이저 조합을 실험하여 최적의 성능을 찾아야 함

F1 기준으로 베스트 모델 선택

동일 모델에서 loss를 기준으로 최적모델을 선택한 것과
validation의 f1 score를 기준으로 최적 모델을 선택한 결과를 비교

Data 증강방식과 베이스라인 모델 학습

Dada 증강방식 적용 유무에 따른 모델 학습 비교

전처리 방식을 달리 하여 학습 진행

줄바꿈 표시 유무와 불용어처리 유무에 대한 전처리 방식을 다르게 진행

외부 데이터를 사용한 학습 진행

일반대화에 대해 외부데이터를 통해 학습진행 시도

모델과 토크나이저 적용 실험 (loss 기준 best model 선택)

모델	토크나이저	f1 score
klue/bert-base	klue/bert-base	**0.9041**
	klue/RoBERTa	**0.8884**
	KoBERT (SKT)	**0.8287**
monologg/koelectra-base-v3-discriminator	monologg/koelectra-base-v3-discriminator	**0.91**
	KoBERT (SKT)	**0.8510**
	klue/RoBERTa	**호환실패**
klue/RoBERTa	klue/RoBERTa	**0.9108**
KoBERT (SKT)	KoBERT (SKT)	**0.7510**

전처리 방식을 달리 하여 학습 진행

모델	전처리 방식	f1 score
monologg/koelectra-base-v3-discriminator	줄 바꿈 표시 X, 불용어 X	0.8977
	줄 바꿈 표시 O, 불용어 X	0.8659
	줄 바꿈 표시 X, 불용어 O	0.8916
	줄 바꿈 표시 O, 불용어 O	0.9010

Data 증강방식과 베이스라인 모델 학습

모델	토크나이저	증강	f1 score
klue/bert-base	klue/bert-base	KorEDA	0.8836
monologg/koelectra-base-v3-discriminator	monologg/koelectra-base-v3-discriminator	KorEDA	0.8987
klue/bert-base	klue/bert-base	GPT-4.1-mini	0.8920
monologg/koelectra-base-v3-discriminator	monologg/koelectra-base-v3-discriminator	GPT-4.1-mini	0.9093

외부 데이터를 사용한 학습 진행

모델	기준	f1 score
klue/bert-base	f1	0.8997
monologg/koelectra-base-v3-discriminator	f1	0.9049

F1 기준으로 베스트 모델 선택

모델 학습 시에는 일반적으로 손실 함수(loss) 값을 최소화하는 방향으로 파라미터가 최적화된다.
그러나 분류 문제에서는 단순한 정확도나 손실 값만으로는 모델의 성능을 충분히 평가하기 어렵다.

따라서 f1-score와 같은 정밀도(Precision)와 재현율(Recall)을 함께 고려하는 종합 평가 지표를 기준으로 베스트 모델을 선정하는 것이 더욱 합리적이다.

본 실험에서는 load_best_model_at_end=True 설정을 통해 검증 과정 중 가장 높은 f1-score를 기록한 시점의 모델을 자동으로 저장하고 불러오도록 하였으며, metric_for_best_model='f1'로 지정함으로써 f1-score를 모델 선택의 기준 지표로 활용하였다.
이로써 단순히 손실이 낮은 모델이 아닌, 실제 분류 성능이 우수한 모델을 선택할 수 있도록 설정하였다.

동일 모델에서 loss를 기준으로 최적모델을 선택한 것과,
validation의 f1 score를 기준으로 최적 모델을 선택한 결과를 비교한 결과
후자의 방식이 실제 테스트 데이터에서의 분류 성능 측면에서 더 일관되고 안정적인 결과를 보였다.

F1 기준으로 베스트 모델 선택 – 베이스라인

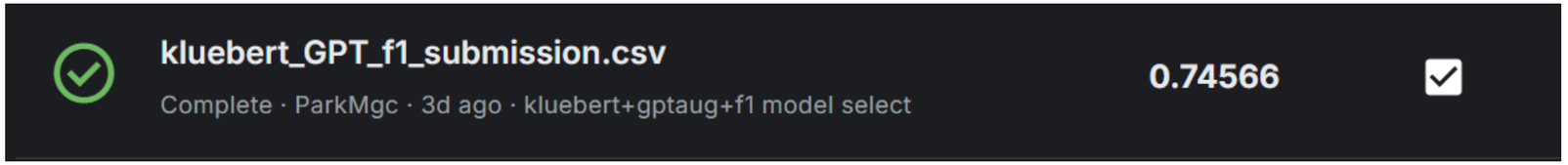
모델	토크나이저	f1 score
klue/bert-base	klue/bert-base	**0.8997**
monologg/koelectra-base-v3-discriminator	monologg/koelectra-base-v3-discriminator	**0.9049**

결과

제출 최종 모델

GPT – 4.1 Augmentation data
+ 일반 전처리
+ klue/bert-base 모델, 토크나이저 사용
+ F1 기반 Best model 선정, Earlystopping

전처리
기본:
- 영어x(원래 데이터에 포함 안됨), 특수기호x(줄바꿈 포함)
- 중복 문자 반복 정규화 (하하하하 -> 하하)
- 데이터 중복 확인 후 정리 (데이터셋 구성 & 전처리 시)



결과분석

- 일반 대화 데이터의 생성과정에서 노이즈 데이터, 혹은 일반적이지 않거나 편향된 분포의 데이터를 형성했을 가능성
- 최소한의 불용어 제거

- Pretrained model을 잘 선택하는 것이 finetuning 성능에 차이를 줄 수 있음
- 부족한 데이터를 증강하는 것이 성능 개선에 도움이 됨
- 한국어에 능숙한 LLM을 활용한 데이터 증강 방식이 한국어 데이터 증강에 도움됨 (KorEDA보다)

Q&A

질의 응답

thank you!