Aiffel research 2N2C

2N
2C

DLThon – Korean Threatening Conversations Classification

김인수 오창원 이동건 홍예린

Aiffel research 2N2C

목차

진행 과정

1일차

EDA

간단한 일반 대화 생성

모델 리서치 및 실험

2일차

복잡한 일반 대화 비교군 생성

모델 실험 및 비교

3일차

데이터 전처리

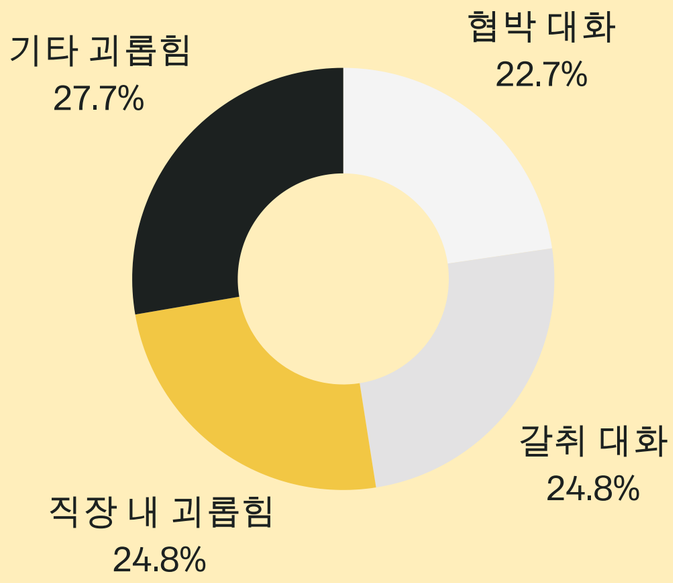
데이터 증강

모델 성능 개선

진행 과정

Exploratory Data Analysis

클래스 분포 확인



학습 데이터에 사용된 특수 기호 목록

{'₩n', ' ', '!', '"', '.', '?', '', '₩u3000', '?'}

테스트 데이터에 사용된 특수 기호 목록

{'₩n', ' ', '!', '"', ',', '.', '?'}

제발 지금 진짜
우리 사람 그래

협박 대화

무슨 그냥 죽여
죄송합니다 아니 그럼

진짜 그럼 아니
지금 그래 없어

갈취 대화

내놔 그냥 제발
무슨 없어요 우리

죄송합니다 대리 회사
아니 지금 오늘

직장 내 괴롭힘 대화

부장 무슨 그럼
사람 우리 그래

진짜 아니 그래
우리 그냥 무슨

기타 괴롭힘 대화

지금 그렇게 아니야
사람 죄송합니다 너무

일반 대화 생성 (feat. chatGPT, solar)

테스트 데이터의 일반 대화

테스트 데이터의 일반 대화는 어떤 것들이 있을까?
- 생성에 앞서, 챗지피티에게 1차 분류를 요청하여 내용 확인

프롬프트 엔지니어링

일단 기본적인 일반 대화를 생성해보자.
- 말을 잘 안 듣네. 잘 듣도록 프롬프트를 만들기

일반 대화의 뉘앙스 ✓

비속어나 가벼운 언쟁, 제 3자 험담 등이 담긴 일반 대화는 어떻게 분류하는게 맞을까?
- 10대 대화, 연인 간 대화 등 주제를 각각 나눠서 생성

모델 리서치 및 실험

모델 선정 기준

1. **한국어** 사전 학습 여부
2. 사전 학습 데이터에 **비속어** 등 **정제되지 않은 데이터**의 사용 여부
3. 사전 학습때 사용한 **전처리 기법**을 모델 학습 때에도 사용 가능
4. **최신** 데이터 셋 사용

GPT2

submission.csv
Complete · Yerin Hong · 1d ago
0.57790

KcBERT

submission_kcvert_test1.csv
Complete · blueZoo · 1d ago
0.66130

KoELECTRA

```
test_data['target'].value_counts()
```

```
3    160
1    125
2    107
0     98
4     10
```

데이터 증강 및 성능 향상을 위한 가설

일반 대화 분류 모호성

- 비속어나 가벼운 언쟁, 제 3자 험담 등이 담긴 대화를 생성

- Test 데이터 A : 기타 괴롭힘으로 라벨링
- Test 데이터 B : 일반 대화로 분류

EPOCH과 일반화

- Early stopping을 사용하여 학습한 모델의 일반화 성능

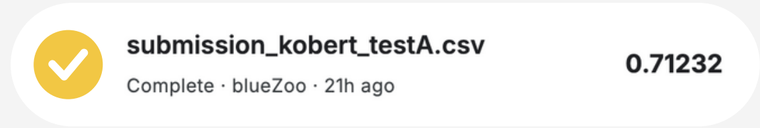
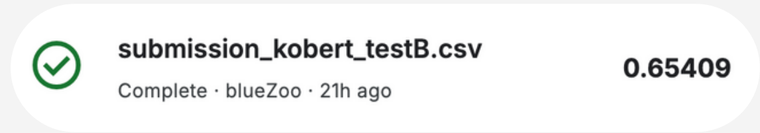
- Early stopping 전, epoch 수를 사용하여 학습한 모델의 일반화 성능

생성된 데이터의 품질

- 무조건 많은 양의 데이터

- 양 보다는 질?

실험 및 베이스라인 선정

일반 대화 분류 모호성 실험	비속어가 속한 대화를 **기타 괴롭힘**으로 분류한 경우			
	비속어 속한 대화를 **일반 대화**로 분류한 경우			
EPOCH 수와 일반화 성능	Early stopping을 사용해 학습한 KcBert와 KoElectra	일반 대화	12	6
	Early stopping 전 epoch 수를 사용한 KcBert와 KoElectra	일반 대화	15	10
생성된 데이터의 추가	Early stopping을 사용한 학습한 KcBert와 KoElectra			
	Early stopping 전 epoch 수를 사용한 KcBert와 KoElectra			

KcBert ← Baseline pick!

- **일반 대화 데이터 추가**

원본 + 일반 대화 데이터로 분류 모델 훈련

- **모델 구조 변경**

분류 모델의 classification head 변경

- **하이퍼 파라미터 조절**

일반화 성능 향상 방법

데이터 전처리

특수 기호 제거

개행 문자 ₩n와 …

토크나이저 개선

- 고유 명사, 비속어, 직급 등 토크나이저 사전에 추가
- 이대리 : 이## ##대## ##리 ##대리, 이대리, 이## 김## 등

Test F1 score

전처리 전후 → 0.71480로 향상

하이퍼파라미터 조절

노드 수

[64, 128, 256]

에폭

[1, 2, 5, 10]

드랍아웃 비율

[0.1, 0.2, 0.3, 0.4, 0.5]

손실함수

[SparseCategoricalCrossEntropy, Focal loss]

모델 성능 분석

검증 F1 score

높을수록 좋음

일반 데이터 분류 개수

클래스의 라벨 개수가 균일할수록 좋음

일반 데이터 분류 정확도

일반 데이터 분류가 정확할수록 좋음

모델 성능 향상

사람도 헷갈리는 어려운 샘플을 분류하기가 어려워요!

Focal loss

모델이 잘 분류하기 어려운 샘플에는 loss를 더 크게 해서 어려운 샘플을 잘 분류하는 데 도움

모델이 데이터셋의 어휘를 몰라요

전처리

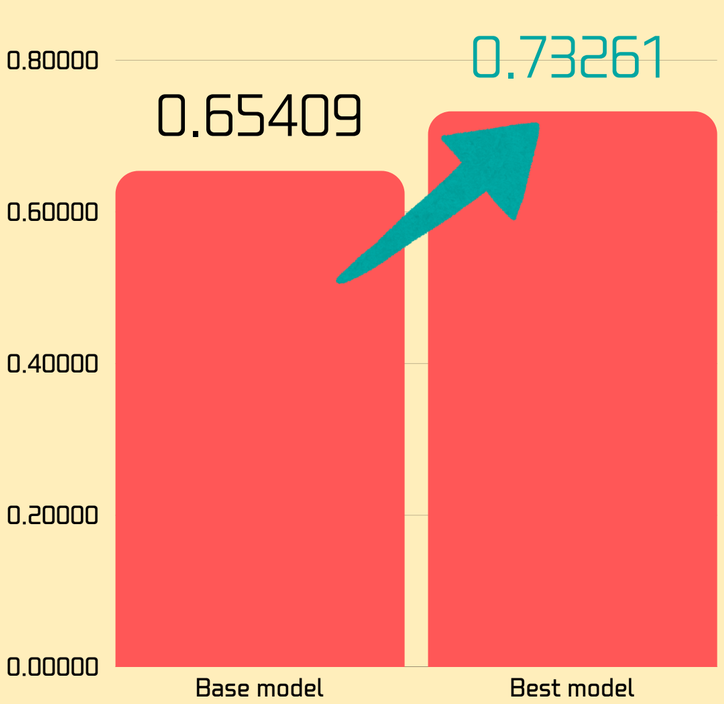
토크나이저에 없었던 어휘를 추가

주어진 시간(=4일, 제출 16회) 내에 모든 모델 공간을 탐색하기 어려워요!

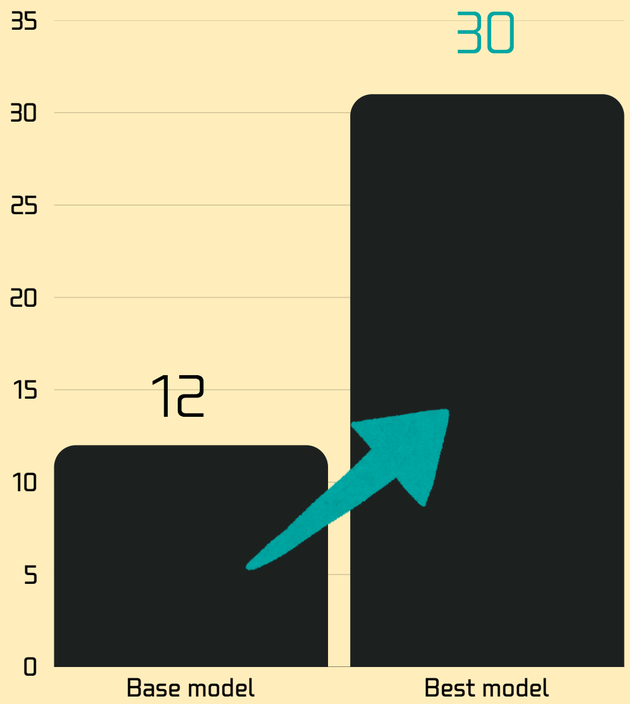
모델군 중 적절한 모델 선택

주어진 시간 내에 튜닝할 수 있는 우수 모델군 선정, 문제 해결에 도움이 되는 하이퍼파라미터 후보군을 서칭

실험 결과



서브미션 점수



일상대화 예측 개수

기준 모델 대비 성능 향상

진행 요약

new_words = ['구급차', '지혈', '침입', '고속도로', '엘리베이터', '진정', '한강공원', '빌라', '소방차','대피', '편의점', '부엌', '심폐소생술', '침착', '출혈', '빌딩', '중학교', '전여친', '전남친', '안부', '지각', '팀장', '##팀장', '팀원', '##팀원', '스캔들', '읽씹', '간식', '프로젝트', '믹스커피', '군것질', '필수품', '힐링', '시발', '씨발', '상사', '피드백', '성의', '레스토랑', '칼퇴', '서운', '퇴사','숙제', '농구', '반장', '쌍욕', '망각', '수다', '떨기', '음료수', '벌점', '지각', '교복', 'PPT', '타이트', '미팅', '희정', '영지', '네일', '파스텔', '자릿세', '보이스피싱', '아이패드', '뺀질', '휠체어', '옥상', '노약자', '##석', '미스##', '데이트', '민아', '말년', '##과장', '과장', '민서', '##씨', '경마장', '숙제', '##대리', '대리', '지수', '교무실', '##주임', '주임', '진희', '병신', '차장', '##차장, '사원', '##사원', '이##', '김##', '박##']
지역 = ['가평', '강남', '거창', '관악', '구로', '나주', '대구', '대만', '대전', '독도', '독일', '동해', '로마', '몽골', '미국', '부산', '부여', '북촌', '북한']
이름 = ['강민', '관우', '광수', '나은', '민재', '민정', '민지', '민철', '민하', '민호', '민희']
비속어 = ['개독', '꼰대', '기레기']
상품/기업명 = ['구글', '구찌', '농협', '던힐', '레종', '맥북', '샤넬', '힐튼', '나이키', '넥서스', '롤렉스', '리니지', '말보로', '빼빼로', '생로랑', '샤오미']

데이터 EDA와 전처리

클래스별 자주 등장하는 어휘를 분석함
데이터의 품사 태깅을 진행후, 고유명사를 필터링하고 토크나이징에 활용
데이터에 자주 나오는 단어들을 모델이 이해하기 쉽게 하나로 묶고 UNK 토큰 등을 완벽히 제거

- KcBert
- KcElectra
- KoElectra
- KoBert
- GPT2

모델군 선정 및 비교 실험

분류 Task를 위하여 사전학습된 2020년대 이후의 한국어 모델

채점 기준 유추

욕설을 포함한 대화를 일반 대화로 볼 수 있을지 A/B test, 욕설 없는 일반 대화가 더 모델 성능에 도움이 되는 것으로 판단

성능 향상을 위해 모델 및 하이퍼파라미터 서칭 공간 정의 탐색 전략 수립

모델의 일반 대화 분류 성능과 전체 성능을 균형있게 향상시키기 위해 전략적으로 베이스라인 모델 및 하이퍼파라미터 서칭 공간을 정의

Aiffel research 2N2C

다양한 모델 및 모델 하이퍼파라미터 서칭

Model\setting	lr	loss=focal	preprocessing(unk제거, 토크나이저에 단어 추가)	pooling	nodes	dropout	에포크	일반대화 분류 개수	추가데이터(solar)+++	train accuracy	val accuracy	f1 score	submission score
kcbert_testB_focal		o	o					2x	o				0.73261
					512/256	0.3	5	35	o	0.9259	0.8672	0.8658	
kcbert_testB_ep1_simple	2e-5		o		x	0.5	1	25	o	0.8382	0.8916	0.8910	
kcbert_testB_ep1_dr4	2e-5		o		256	0.5/0.5	2	10	o	0.9261	0.9160	0.9161	
kcbert_testB_ep1_dr4	2e-5		o		256	0.5/0.5	1	tp 30(out of total 33)	o	0.7673	0.9097	0.9086	
kcbert_testB_ep1_dr3	2e-5		o		128	0.3	1	25	o				0.71480
kcbert_testB_ep5_dr3	2e-5		o		128	0.3	2	15	o				
kcbert_testB_ep2_dr3	2e-5		o		128	0.3	1	9	x				
kcbert_testB_ep2_dr3	2e-5		o		128	0.3	2	4	x				
kcbert_testB_ep2_dr3(changwon)	2e-5		x		128	0.3	2	16	o				0.70714
kcbert_testB_ep2_dr3	2e-5		x		128	0.3	2	40	x				
kcbert_testB_ep2	2e-5		x		128	0.1	2	15	x				
kcbert_testB	2e-5		x		128	0.1	4	12	x	0.9804	0.8884		0.65409
kcbert_testA	2e-5		x		128	0.1	5	22		0.9912	0.8873		0.71232
kcbert								-					0.66130
kcElectra	2e-5		x			0.1	8	6	x				-
kcElectra	2e-5		x			0.1	5	15	o				
kcElectra	2e-5		x			0.3	5	27	o				-
kcElectra	2e-5		x			0.5	5	11	o				-
kcElectra	2e-5		o				5		o				-
gpt2													

훈련 정확도,
검증 F1 score,
일반 데이터 분류 개수,
일반 데이터 분류 정확도를 평가 지표로
사용

과적합 문제
Early stopping, Dropout 비율 조정 및 데이터
추가 통하여 과적합에 대응

일반 대화 분류 성능,
과소적합과 과적합을 나타내는 지표를
종합적으로 판단하여
모두 우수한 모델 선택하여 제출

Q&A 세션

경청해주셔서 감사합니다!