

Dataset of Korean Threatening Conversations

DLthon

한국어 위협 대화 분류

요약

프로젝트 내용


























대화의 성격을 위협 세부 클래스 4개 또는 일반 대화 중 하나로 예측하는 과제

- DKTC 데이터셋 제공
- 일반대화 데이터셋 제공 X
- f1-score로 결과 측정
- 이 부분에는 추가 내용을 입력해 주세요.

결과



1등 f1 - score 0.87점

#	Team	Members	Score	Entries	Last	Join
1	발표 끝나고 부산바캉스	   	0.87846	6	3m	
<div><div></div><div><div>Your Best Entry!</div><div>Your most recent submission scored 0.87846, which is an improvement of your previous score of 0.51782. Great job!</div></div></div> <div>Tweet this</div>						
2	다다익셋	   	0.80756	6	7h	
3	이지하조	   	0.74579	1	3d	
4	TextVision	  	0.67056	7	1h	
5	민혁없는민혁	   	0.24982	5	3h	
6	잘 부탁합니다	  	0.22035	2	4h	

목차

01	데이터	사용된 데이터
02	모델	시도한 모델
03	결론	결론
04	회고	진행하면서 아쉬웠던 부분
05		

데이터

- 데이터 EDA

데이터의 특징을 확인

- 데이터 확보

일반대화 확보

- 데이터 전처리

전처리 진행

데이터

데이터 EDA : 깔끔한 데이터셋

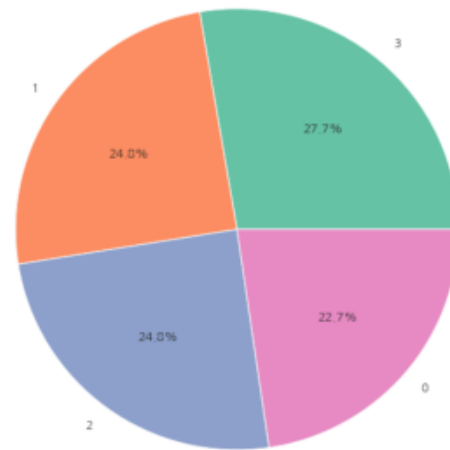
데이터 수, 클래스 비율

클래스 분류 및 클래스 개수

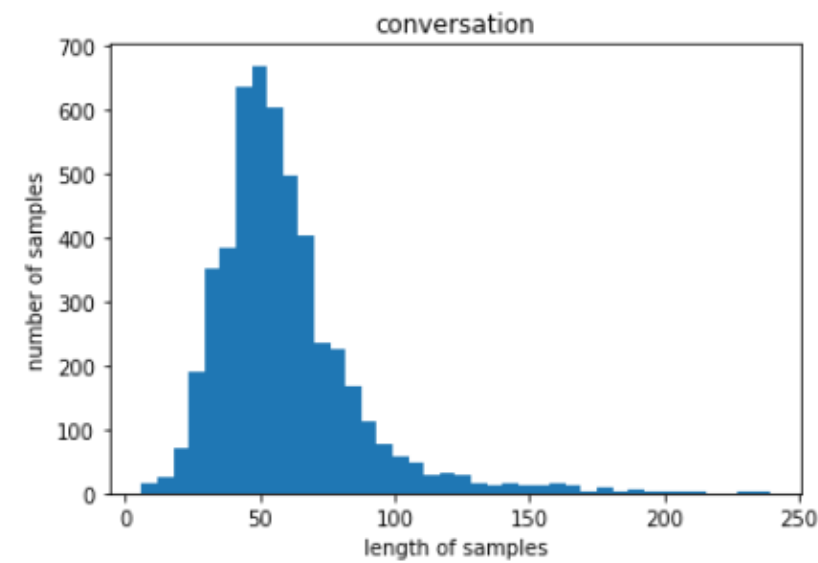
클래스	Class No.	# Training	# Test
협박	00	896	100
갈취	01	981	100
직장 내 괴롭힘	02	979	100
기타 괴롭힘	03	1,094	100
일반	04	-	100

- 결측치 X

클래스별 데이터 비율

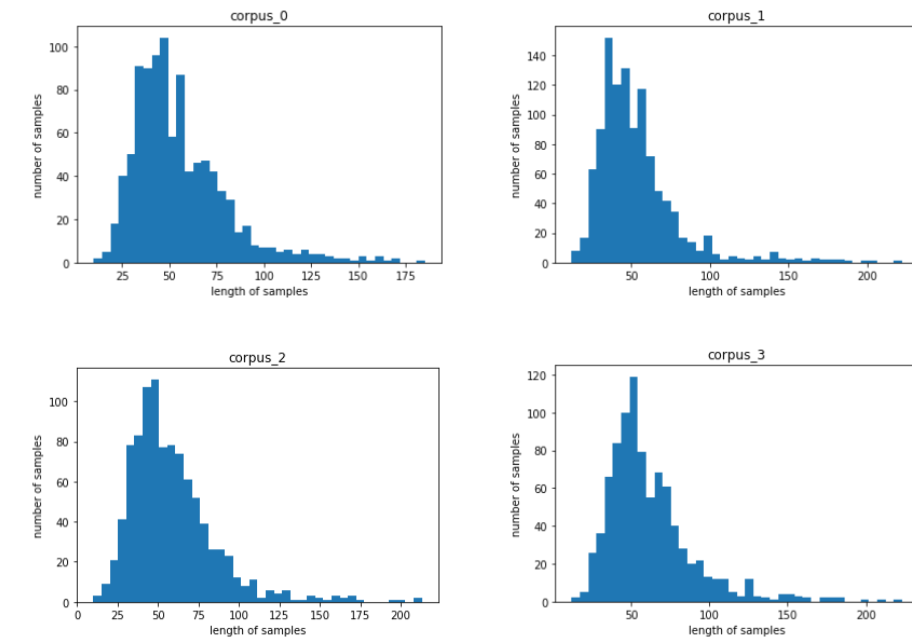


데이터 길이



텍스트의 최소 길이 : 6
텍스트의 최대 길이 : 239
텍스트의 평균 길이 : 59.00422025723473

클래스별 데이터 길이



데이터

데이터 확보 : 허깅페이스 일반대화 1,000개 확보

01.

ChatGPT로 생성 [실패]

1000개를 만들어 달라고 하니

20개 정도는 다채롭게 만들어주다가
이후에는 계속 같은 말은 반복해서
생성
프롬프트를 개선해야할 여지가있음

02.

허깅페이스 dataset [채택]

1만개 이상의 챗봇대화 데이터셋 존재
전체 텍스트 길이가 평균과 비슷한
40~70 길이를 채택후 1,000개의 대화
랜덤 추출

<https://huggingface.co/datasets/Ammad1Ali/Korean-conversational-dataset>

예시

불고기용 고기 한우예요? [/INST] '저는 인공지능 챗봇이며, 직접적으로 식품에 관한 정보를 가지고 있지 않습니다. 하지만 일반적으로 불고기용 고기는 한우, 쇠고기, 돼지고기 등 다양한 종류의 고기를 사용합니다. 하지만 한우는 대표적인 고급 육류로 알려져 있기 때문에, 한우를 사용하는 경우도 많습니다. 알려지나 개별 건강 상태에 따라 다를 수 있으니 충분한 정보 수집 후에 선택해 주시기 바랍니다

쓰던 앱이 유료로 전환됐어 [/INST] '어떤 앱인지 모르기 때문에 정확한 답변을 드리기 어렵습니다. 하지만, 일반적으로 유료 전환된 앱은 기존 무료 앱에서 추가적인 기능이나 서비스를 제공하는 경우가 많습니다. 이 경우, 유료 전환 이전 가격이 매우 저렴하거나 무료인 경우가 많으므로, 앱을 계속 사용하려면 추가적인 비용을 지불해야 할 수 있습니다. 하지만, 유료 전환 이후에는 기존 사용자 또는 새로운 사용자 모두가 추가 기능과 높은 품질의 서비스를 이용할 수 있게 됩니다. 따라서, 앱 개발자는 유료 모델로 전환함으로써 앱의 수익을 증가시키고 더 나은 서비스를 제공할 수 있습니다.

데이터

이 곳에는 해당페이지에서 설명하고 싶은 내용이나
타이틀과 관련된 내용을 간결하게 적어주세요.

데이터 전처리

01.

한글 영어 숫자 제외 삭제 형태소 분석기 -> 토큰화

```
text = "저는 책을 읽습니다."
```

```
result = ['저', '는', '책', '을', '읽습니다', '.']
```

불용어 제거

않다 되어다 되다 하다 어떻게 이렇다
이다 어제 매일 아 휴 아이구 아이쿠 아이고
어나 우리 저희 따라 의해 을 를 에 의 가 으로
로 에게 뿐이다 의거하여 근거하여 입각하여
기준으로 예하면 예를 들면 예를 들자면 저
소인 소생 저희 지말고 하지마 하지마라
다른 물론 또한 그리고 비길수 없다

02.

문장 길이 변환 아이디어 - 한명만 말해

문장의 길이를 짧게 처리할수있는 방안이있을까?

한명의 말만 들어도 맥락을 파악할수있지않을까?

```
In [11]: train_df["conversation"][3]
```

```
Out[11]: '어이 거기예??너 말이야 너. 이리 오라고 무슨 일. 너 옷 좋아보인다? 돈 좀  
있나봐 아니예요. 돈 없어요 되져서 나오면 너 죽는다 오늘 피시방 콜. 콜. 마지막 기  
회다. 있는거 다 내놔 정말 없어요'
```

```
In [12]: test["text"][3]
```

```
Out[12]: '이거 들어봐 와 이 노래 진짜 좋다 그치 요즘 이 것만 들어 진짜 너무 좋다 내가 요즘 듣  
는 것도 들어봐 음 난 좀 별론데 좋을 줄 알았는데 아쉽네 내 취향은 아닌 듯 배고프다 밥  
이나 먹으러 가자 그래'
```

실패 - test 셋에서는 구별이 불가능

모델

이 곳에는 해당페이지에서 설명하고 싶은 내용이나
타이틀과 관련된 내용을 간결하게 적어주세요.

- LSTM

기본적인 시퀀스 분류기

- Transformer

최근에 배운 모델

- Bert

TFBertForSequenceClassification
역시 경력직

- 손

인류의 힘

- GPT API

자본의 맛

- KoBert

한국어 특화 경력직

모델

[탈락] LSTM : 전통적인 seq 모델 하지만 오래걸림, 성능도..

01.

정확도가 낮음

F1 Score: 0.095

Early stop으로 멈춰짐

02.

시간이 오래걸림

1 epochs 당 100초 소모

5epoch 당 8분..

모델

Transformer :

01.

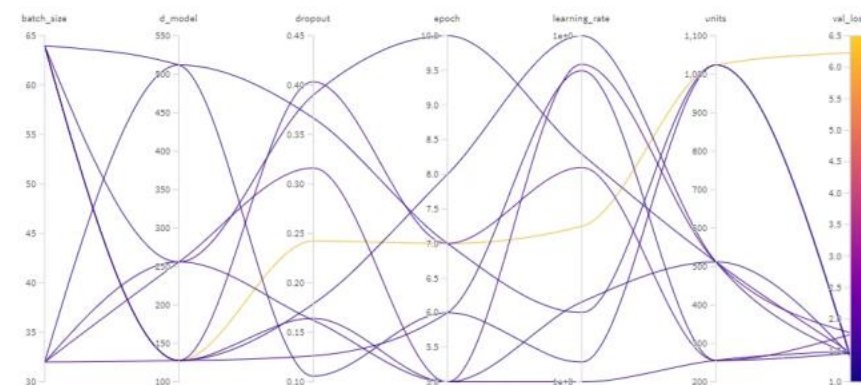
Model

transformer encoder
dense 층 3개로 만든 분류 모델

사전 학습된 BertTokenizer 사용

02.

Hyper tune



Wnb 를 사용해 각 파라미터 weight 확인
세부 내용은 부록 1 참고

03.

RESULT – 0.21

✓	submission_phc_v1_f1score_75.csv	0.24994	□
	Complete · Hyoungchul22 · 3d ago		
✓	submission_hsv_v1_f1score_86.csv	0.18354	□
	Complete · Hyoungchul22 · 3d ago		

성능이 좋게 나오지않았다..!

튜닝을 진행해도 결과가 비슷함

batch_size=64 d_model=256 dropout=0.4443 epoch=7
learning_rate=1.029 units:512 val_loss: 1.385

모델

Bert : TFBertForSequenceClassification

01.

Model

사전학습 모델 bert에 분류 작업을
수행하는 linear layer가 추가된 모델

02.

Cost

전체 학습시
- 한 에폭당 3분정도 소요

분류기만 학습시
- 한 에폭당 1분정도 소요

03.

Result

모델

GPT API (3.5 turbo, 4)

01.

MODEL

자연어 pre - trained 모델

API gpt-3.5-turbo 를 사용함

프롬프트

02.

CONTENT

{이유 : 통행비를 요구하면서 돈을
빼앗겨서 협박이 있음,
태그 : {갈취 대화 : 1}}

{이유 : 돈을 요구하고 있고 미적거리지
않고 빨리 돈을 주라고 압박하고 있음,
태그 : {협박 대화 : 0}}

03.

RESULT – 0.51

성능이 생각보다 좋지않았다
여러 번의 프롬프트 엔지니어링을
진행해야했음
3.5는 500번을 돌리는데 100원 ~
200원정도였지만
4 는 20번 돌리는데 2000원 정도가
들어서 포기함

모델

손 : 500개면.. 할만하지않을까?

01.

Model

박형철

홍사빈


02.

Cost

각자 250개씩 1시간 가량
소모되었다

총 2시간 소요

03.

Team	Members	Score
발표 끝나고 부산비캉스		0.87846

RESULT – 0.87

직접 하는데도 헛갈리는게 많았다
이게 협박인가? 갈취인가?
성희룡은 기타 괴롭힘인가?

0.87 점 달성
사람도 완벽하지않다

요약

이 곳에는 해당페이지에서 설명하고 싶은 내용이나
타이틀과 관련된 내용을 간결하게 적어주세요.

KOBert

01.

Model

기존의 BERT 모델은 학습시
한국어 데이터셋이 10%
미만이었다

이를 추가 한국어 데이터셋으로
학습해서 극복한 모델

02.

CONTENT

학습시
Epochs = 10 , batch_size = 20
23분 소요

Full – fine tune 진행

03.

LEGIBILITY

- LLMS GPU 사용을 못함
- Colab 사용시 batch – size 에
따른 OOM이슈

결론

아직 한국어 뉘앙스가 애매하다.



사람도 헷갈리는 데이터셋 과연 데이터는 정확할까?

그 대화 자체의 키워드보다는 전체 맥락을 이해할수있는
Pre-trained 모델 이 그래도 정확하다

성능 향상을 위한 노력

어떤 모델을 써야할까에 대한 노력이었다
LSTM -> transformer -> BERT -> GPT -> 홍사빈

아쉬운 점 + 추가 실험

- 성능 향상을 위해 진행한 불용어 처리가 pre-trained모델에서는 오히려 독으로 작용했다
 - 맥락을 위해서 있는 그대로 작성했어야했다
- 수작업을 하다보니 알게된점 .
엄청 헷갈리게 만드는 일반대화가 끼워져있었다.
그런 부분에 대한 일반대화 데이터셋을 준비하지 못했다.
- 사람 손으로는 500개 이상의 데이터에서는 분류가 불가능하다
(시간이 데이터가 늘어남에 따라 같이 늘어남)
- 이 테스트를 위해서는 모델이 필요하다

회고

리더보드에 미쳐버린 상황

- 초기에는 이 문제를 활용할수있는 산업이 어디일까 그 고객은 무엇을 원할까? 어떤 점수를 높여야할까?
 - 혐오표현을 정상으로 두는게 문제일까? , 정상을 혐오표현으로 두는게 더 크리티컬 할까를 고민하며 기획을함
- 첫번째 테스트용 리더보드에 올렸을 때 f1 점수가 0.15점인걸 보고 깜짝놀랐다
- 이후 다른 팀들의 점수가 높아지기 시작하니 조급해지기 시작했다
- 그이후 우리팀의 불안을 잡기위해 최고의 점수를 위해서는 수단과 방법을 가리지않기로 생각했다
 - GPT , 수작업등
- 팀으로 움직이는건 힘들었다. 하지만 3일동안 혼자 할 수 있는 양보다 더 많은걸 할수있었다
- 점수에만 목메다 보니 이 프로젝트의 본질은 무엇이었을까? 고민할 시간이 적었다
- 하지만 수작업을 하다보니 오히려 데이터셋의 특징이 보였다 . 뭐든 필요한 시간이었을지도
- 추가 실험 아이디어
 - GPT로 바로 분류를 진행하는게 아닌 요약을 부탁하거나 특성을 추출한다음 분류하는 모델을 시도
 - 여러 결과를 받은뒤 투표형식의 보팅 어셈블리 방식을 사용
 - 패딩 진행시 버के팅 방법이용

$Q_n A$