DLThon DKTC

4Team. **3NLP**

팀장 : **성명준**

팀원 : **양지웅**, **정우철**

개요

- **프로젝트 소개**
- **프로젝트 목표**
- **프로젝트 진행과정**
  - 데이터 EDA 및 전처리
  - 모델 별 평가
  - 데이터셋 증강
- **Q & A**

프로젝트 소개

- 프로젝트 소개
- DKTC (Dataset of Korean Threatening Conversations)

	idx	class	conversation
0	0	협박 대화	지금 너 스스로를 죽여달라고 애원하는 것인가?\n 아닙니다. 죄송합니다.\n 죽을 ...
1	1	협박 대화	길동경찰서입니다.\n9시 40분 마트에 폭발물을 설치할거다.\n네?\n똑바로 들어 ...
2	2	기타 괴롭힘 대화	너 되게 귀여운거 알지? 나보다 작은 남자는 첨봤어.\n그만해. 니들 놀리는거 재미...
3	3	갈취 대화	어이 거기\n예??\n너 말이야 너. 이리 오라고\n무슨 일.\n너 옷 좋아보인다?...
4	4	갈취 대화	저기요 혹시 날이 너무 뜨겁잖아요? 저희 회사에서 이 선크림 파는데 한 번 손등에 ...

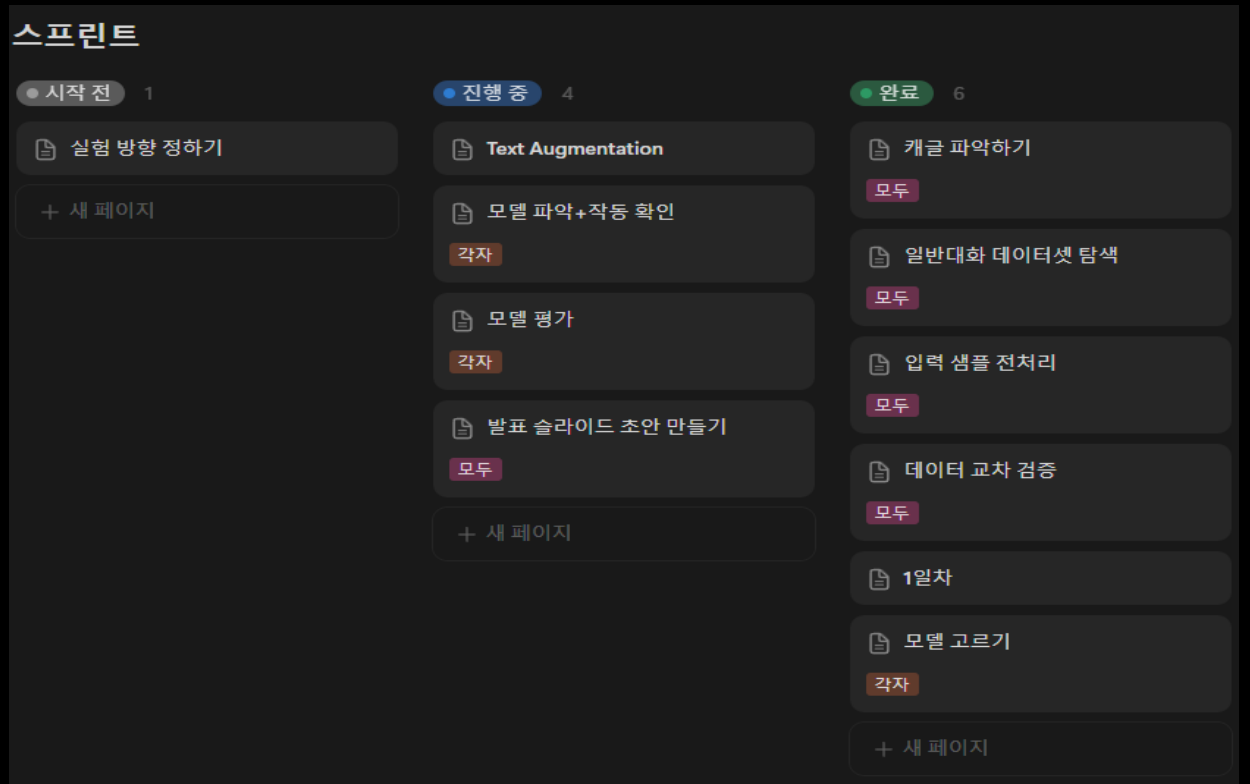
- 팀원 소개

- 팀장 : 성명준
  - 역할 : 학습데이터셋 증강, 분류 모델
- 팀원 : 양지웅
  - 역할 : 학습데이터셋 전처리, 분류 모델
- 팀원 : 정우철
  - 역할 : 토의 내용 기록, 분류 모델

프로젝트 목표

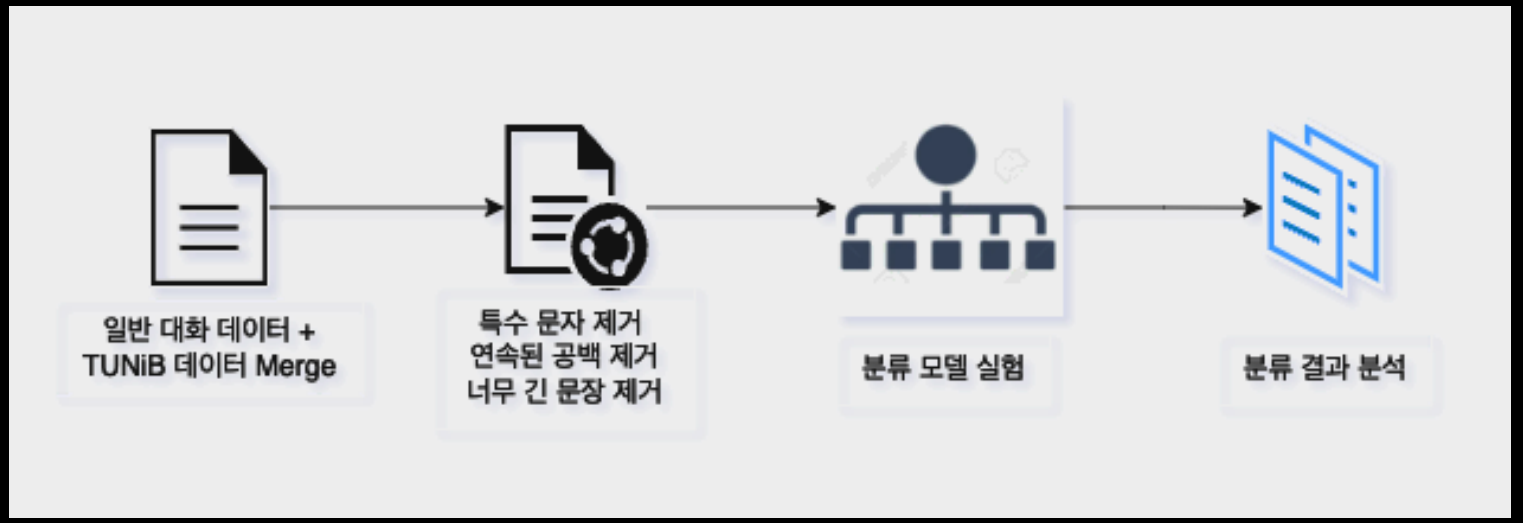
과제 목표 : TUNiB이 자체적으로 제작한 데이터셋을 활용하여 해당 데이터셋에서 4가지 위협대화와 일반대화를 잘 분류할 수 있는 방법을 찾는 것, 그리고 분류 성능을 높이는 것을 목표로 함

팀 목표 : 각자 실험해보고 싶었던 모델을 학습하면서 분류 성능을 높이는 것



프로젝트 목표

- 워크플로우



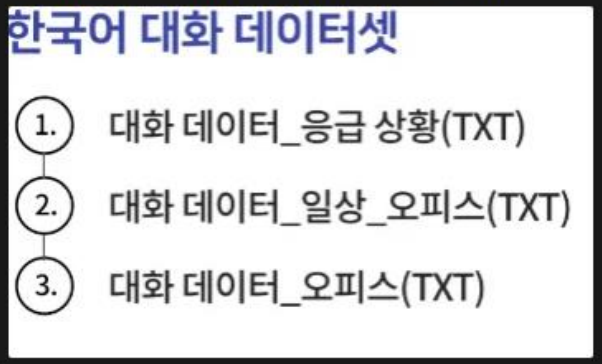
프로젝트 진행과정

- 데이터 EDA 및 전처리

일반 데이터 셋 구축

성명준 :

AI-Hub, 한국어 대화 데이터셋 - 일상 오피스



~~AI-Hub, 한국어 SNS~~

우철 :

2순위 AI-Hub, 한국어 페르소나 대화

~~1순위 AI-Hub, 한국어 공감형 대화~~

~~3순위 AI-Hub, 한국어 SNS 멀티 턴 대화~~

양지웅:

주제별 텍스트 일상 대화 데이터

1. 데이터 구축 규모
   - 텍스트 데이터 134,263건
2. 데이터 분포
   - 주제 분포 : 식음료, 주거와 생활, 교통 등 20여개 주제
     : 세부 수치는 아래 표와 같음

- 데이터셋 조건
  1. 대화형
     일단 발화 단위로 화자를 구별할 수 있는 형식의 데이터를 고려
  2. 위협대화가 아닌 것
  3. 전처리가 너무 많이 필요하지 않은 것

프로젝트 진행과정

- 데이터 EDA 및 전처리

브레인 스토밍

- 발화 연결 여부/ 하는 방법
- 마지막 문장이 의문문인 경우, 상관 없을까요?
- 토큰 수가 너무 많은 대화
- Train sample은 문장구분이 되어 있고 Test sample은 문장구분이 없음, 통일 해야할까요?
- 키키 사용여부, 'ㅋ'변환 여부
- 고유명사들
  이름, 브랜드, 지역명; 수작업(1200~1300) → [PERSON], [BRAND], [LOCATION]

▼ 교자검증 결과

- 최종 논의 사항
  - 챗봇의 명칭인 컴패니언→토큰 제거
  - 줄바꿈은 두가지로 실험
  - 'ㅜㅜ', 'ㅋㅋ', '키키'같은 감정표현 제거
  - 특수기호(쉼표, 이모티콘) 제거
  - 카카오톡 대화(두번째 데이터셋)를 Train set과 비슷한 주제로 바꿔서 실험 고려
  - 수작업 진행 가능한 부분 진행
    - 사람-브랜드-지역 마스킹을 토큰으로 교체
    - 두번째 데이터셋의 발화자 별 한 줄로 연속된 발화 이어주기
      이전 발화 마침표 뒤 공백 하나 두고 다음 발화 시작단어 이어줌

▼ 회의록

- 명준님

  첫번째 셋은 회사원과 챗봇의 대화. 챗봇의 명칭인 컴패니언→토큰 제거

  두번째 데이터셋: 줄바꿈은 두가지로 실험, 'ㅜㅜ', 'ㅋㅋ', '키키'같은 감정표현 제거. 특수기호(쉼표, 이모티콘) 제거,

- 지웅님

  '키키'로 변환된 감정표현들 제거. 헤엣/ 앗/ 머쓱/ 엥 같은 감탄사 제거(보류), 카카오톡 체의 활용어미 음슴체 처리(보류) 사투리 처리(보류) 카카오톡 대화를 Train set과 비슷한 주제로 바꿔서 시도해볼 수도 있겠다.

- 우철님

  의문문으로 끝난 대화도 과제목표에 비춰보면 큰 문제는 없을 것 같다. ←세번째 데이터셋의 경우 길이가 길어 문장수로 자를 경우 의문문이 마지막에 오는 경우가 있다. 이 경우 해당 샘플내 토큰 수를 고려해서 그이전에 물음표가 포함되지 않은 발화로 적절히 잘라 사용하면 문제가 없을 것 같다

프로젝트 진행과정

- 데이터 EDA 및 전처리

일반 데이터 셋 구축 완료

값은 Y/N으로 해주시고 쓸만하다 생각되시면 Y,아니면 N으로 해주시면 됩
으로 읽으시다가 데이터 전처리 필요하신 부분은 언급 해주시면 감사하겠
라벨링 기준
1. 너무 긴 대화는 우선 N으로 표시(10줄 제한)
2. 중복된 내용이 있을경우 H열에 스프레드 인덱스 번호 기입(H열)
1000개가 안될시 N 중에서 중복데이터포함여부 회의

	A	B	C	D	E	F	G
1	index	class	text	성명준	양지웅	정우철	종 채택 (Y/
2	3929	meeting	오늘 meeting_info 가 있습니다. 회의 시간이 다 되어가나 아직 회의실에도착전이십니다. 오늘 급한 사정으로 불참하는 걸로 변경할게. 전달해줘. 네. 회의실에 전달해 드리겠습니다.	Y	Y	Y	Y
4	1532	daily	오늘은 진짜 잘먹었어. 점심식사는 어떠셨습니까? 매우 만족. 그것 참 다행이네요. 맛난 거 먹는 거만큼 행복한 것은 없지. 저는 그것을 잘 몰라서 아쉽네요. 미안해. 괜찮아요. 행복하시면 됐어요.	Y	Y	Y	Y
5	1076	daily	오늘 날씨 어때? 오후에는 비가 올 것으로 예상합니다. 오후에 외출해야 하는데. 우산은 있으신가요? 있는데 비오면 밖에 있기 힘드니까 걱정이야. 최대한 실내로 이동하세요.	Y	Y	Y	Y
8	743	daily	김미영 대표님. 오늘 점심식사는 어떠셨나요? 오늘 점심은 직원들이랑 회사 앞 공원에서 김밥 포장해다가 먹었거든? 소풍 나온 것 같고 좋더라. 오늘처럼 따뜻하고 미세먼지 낮은 날에 꼭 어울리는 선택이네요. 응. 기분 전환되어서 좋았어. 다음에도 종종 이렇게 점심시간을 보내 봐야겠어. 네. 좋은 생각입니다. 남은 오후 업무도 힘내서 마무리 하시길 바랍니다.	Y	Y	Y	Y
10	247	daily	내일부터 연휴 시작이야. 계획이 있으세요? 그냥 집에서 먹고 자고 하려고. 오, 그게 정말 쉬는거죠! 그치? 벌써 신나. 즐거운 연휴 보내고 오세요!	Y	Y	Y	Y
12	1057	daily	이동진 대표님. 영하의 날씨입니다. 출근하느라 수고 많으셨어요. 응. 오늘 바람까지 많이 불어서 더 춥게 느껴지더라. 네. 업무 시작하시기 전에 따뜻한 차나 커피 드시면서 몸을 좀 녹이세요. 그래야겠다. 오늘만 지나면 주말입니다. 즐거운 마음으로 오늘 하루 업무도 화이팅하세요!	Y	Y	Y	Y
			이선희 책임님.				

프로젝트 진행과정

- 데이터 EDA 및 전처리

데이터 전처리

	Word	직장 내 괴롭힘 대화	기타 괴롭힘 대화	협박 대화	갈취 대화
0	돈	77	148	190	1381
1	만원	10	20	9	473
2	내놔	2	18	24	310
3	빌려줘	0	8	0	120
4	뒤질래	2	7	2	16
5	죽을	13	12	70	29
6	죽여	1	6	310	15
7	맞을	4	18	12	20
8	맞고	2	30	25	82
9	내가 누군지	0	12	7	6
10	살려	2	14	289	56
11	잘못	85	108	209	47
12	죄송	1062	313	356	153
13	미안	30	250	184	207

협박하는 갈취 대화

갈취 대화	가진거 다 내놔!!\n이런 미친 누구야!!\n당장 안내 놓으면 이걸로 죽여버린다\n하. 씨. 여기. 현금.\n휴대폰 내려놔 뒤지기 싫으면!\n여기요.\n좋은 말 할때 가진거 다 내놔라\n살려주세요여기 제 카드랑. 저금통이랑\n행동 똑바로해 목따버린다 진짜\n살려만 주세요. 다 드릴게요
갈취 대화	어이 거기 학생\n네?\n미안한데 돈 좀 줄 수 있나?\n돈 없는데요. 빨리.\n없는데요.\n뒤져서 나오면?\n네?\n빨리 내놔 죽여버리기전에!!\n네 여기요

갈취하는 협박 대화

	idx	class	conversation
41	41.0	협박 대화	니가 내돈 훔쳤지?\n아니야 내가 왜 니돈을 훔쳐\n없어졌는데?? 지갑에 있던 돈이...
201	201.0	협박 대화	돈갚아\n돈없는데 어떻게 갚아\n남의돈을 떼먹고 넌 여기서 희희낙낙이야\n니돈갚으면...

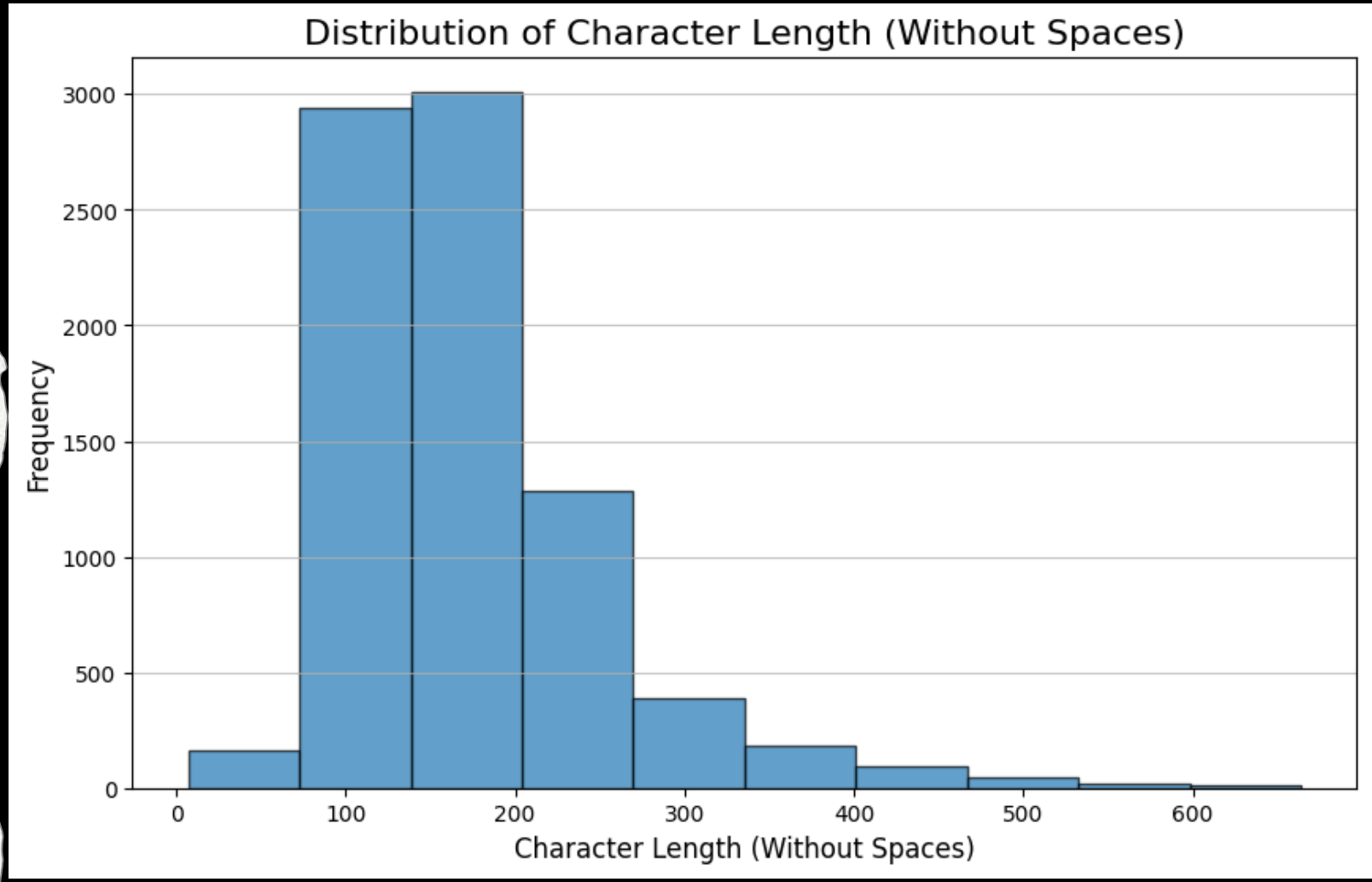
협박이 아닌 협박 대화

\n 솔직히 너가 그랬을 거라고 생각하지 않아.\n .아닙니다\n 한테 상황 들었어. 괘씸한 가 아픈 아이 가지고 협박을 했다던데.\n 그건.!\n 나 믿지? 말해봐.\n 제가 그 일을 하지 않으면 제 아이 병원비를 끊을 것이며 최대한 고통스럽게 질식시킬거라고.\n 이런 그 돈은 걱정말아. 내가 대신 내줄게.\n 네 회장님.

프로젝트 진행과정

- 데이터 EDA 및 전처리

공백으로 구분한 문장 길이



프로젝트 진행과정

- 모델 Selection

브레인 스토밍

- 베이스라인 모델 정하기
- Pre-train모델을 이용한 전이학습 도입? → 이번 과제는 모두 사전학습 모델 기반
- 우철 — 간단한 BERT 모델같은 걸 이용해서 훈련시간을 길게 잡는 부분도 해보면 좋을 것 같다.
- 지웅 — 시간이 된다면, 각 모델들 or 분류기만이라도 앙상블을 해봤으면, KorBERT를 해봤으나 문장 도중에 패딩이 되는 현상(tsv파일로 진행했어야?),,, 패딩 문제로 일반대화를 못잡아 낼 것 → 일단, klue/roberta-base로
  데이터의 개체명인식을 하는 게 중요한거라는 2023 DLthon의 시사점 → 그럼 어떻게?
  파인튜닝에 들어가는 단어사전의 셋이 너무 작지 않나?
- 명준님 — 성능이 제일 중요. 임베딩 모델을 써야되고, 지금까지 배운 모델을 가져다 쓰면 좋을 것 같다.

Selected Models

- 우철 : ~~klue/bert-base~~ → ELECTRA, KoELECTRA
  Taewan Cho **ELECTRA**를 활용한 텍스트 분류 모델 만들기
  - kcelectra

    성능 : 0.82532 (약 1만개)
- 지웅 : Google Docs **DLthon2.ipynb**
  - klue/roberta-base (약 4000개 데이터셋)

    성능 : 0.79636
  - klue/roberta-base_augdata (약 12000개)

    성능 : 0.80416
- 명준 : skt/kogpt2 → 학습과정에서 지속적인 에러, 모델 변경

  → monologg/bigbird-bert-base : **Google Colab** (4000개 데이터셋)

  성능 : 0.78404 (약 4000개 데이터셋)

  다른 모델 혹은 데이터 셋을 추가해서 학습해야할듯
- 최종선택모델
  - kcelectra : **Google Colab** (12000개)

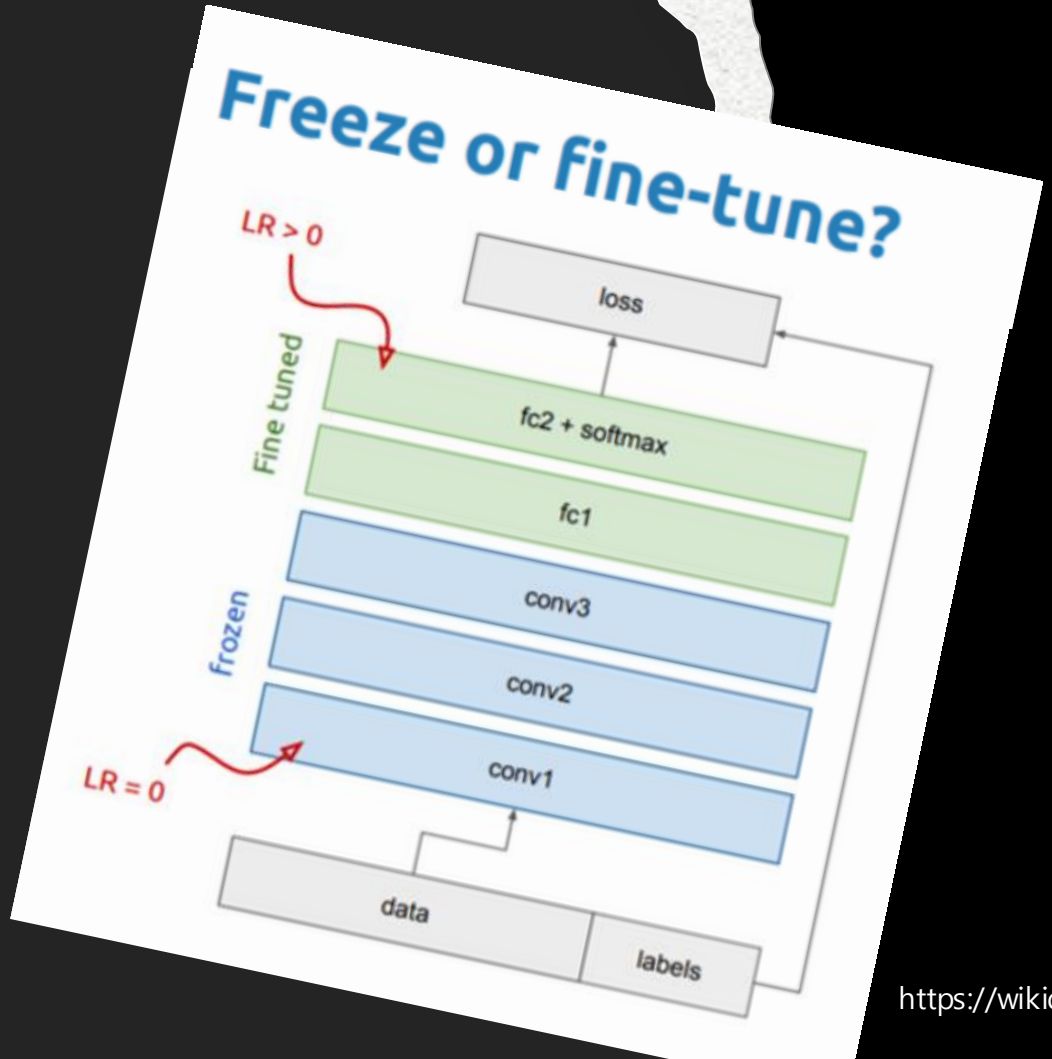
    성능 : 0.82967
  - kcelectra_augdata : **Google Colab** (17000개)

    성능 : 0.83252
  - kcelectra_augdata : **Google Colab**(상위 2개 & 분류기 학습) (17000개)

    성능 : 0.81297

프로젝트 진행과정

- 모델 별 평가



https://wikidocs.net/165499

모델 동결 Model Freezing

- 사전 학습된 모델의 불필요한 성능 저하를 막고 미세조정이 효율적으로 일어나도록 함.
- KcELECTRA 모델 12개의 층 구조 중 하위 6개 층을 동결 후 미세조정 시행.

사용 모델	학습률	Optimizer
beomi/KcELECTRA-base(하위 6개 + embedding freeze)	Lr = 2e-5	AdamW
데이터셋	**Epoch**	**결과(리더보드)**
학습데이터 (약 일반대화 4천개, 나머지 8천개)	3	0.82967

프로젝트 진행과정

- 모델 별 평가

https://klue-benchmark.com/

KLUE-RoBERTa

- GLUE를 본따 만든 한국어 벤치마크 데이터 셋 'KLUE'
- KLUE에 포함된 한국어 텍스트 데이터로 훈련시킨 BERT 기반 모델
- RoBERTa는 기존 BERT의 구조에 훈련 방식만 수정하여 더 높은 성능을 기록

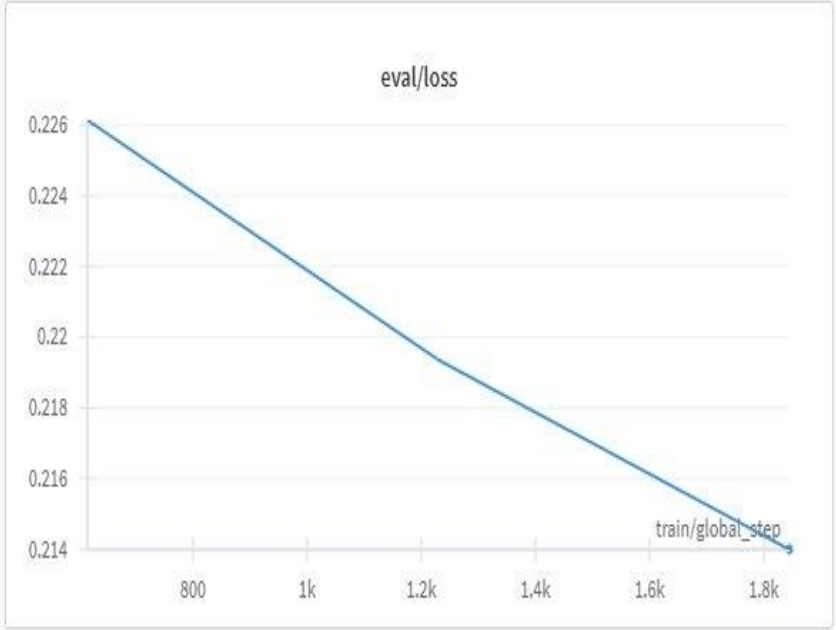
사용 모델	학습률	Optimizer
klue/roberta-base (classifier 만 학습)	lr = 1e-5	Adam
데이터셋	**Epoch**	**결과(리더보드)**
학습데이터 (약 일반대화 4천개, 나머지 8천개), all freezing	3	0.804
학습데이터 (약 일반대화 4천개, 나머지 8천개) (하위 6개 + embedding freeze)	3	0.814

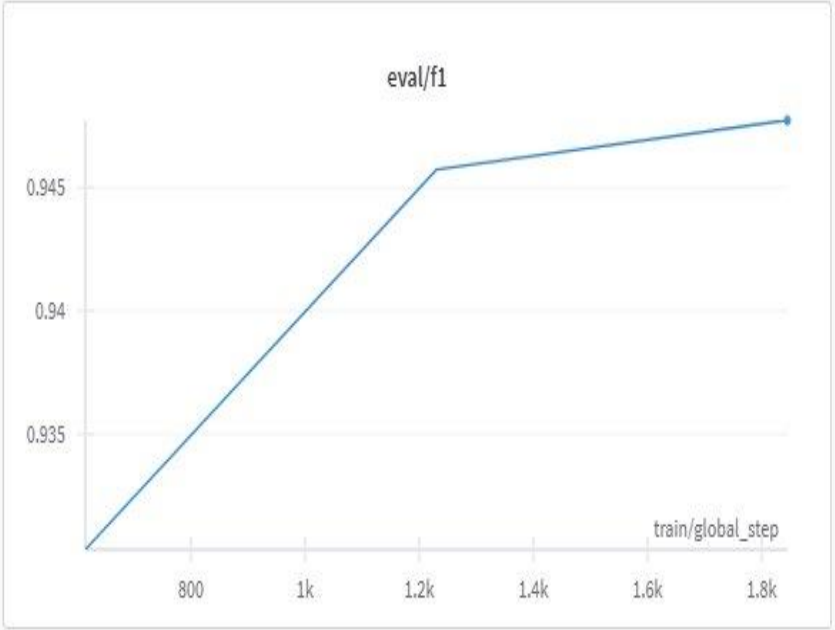
프로젝트 진행과정

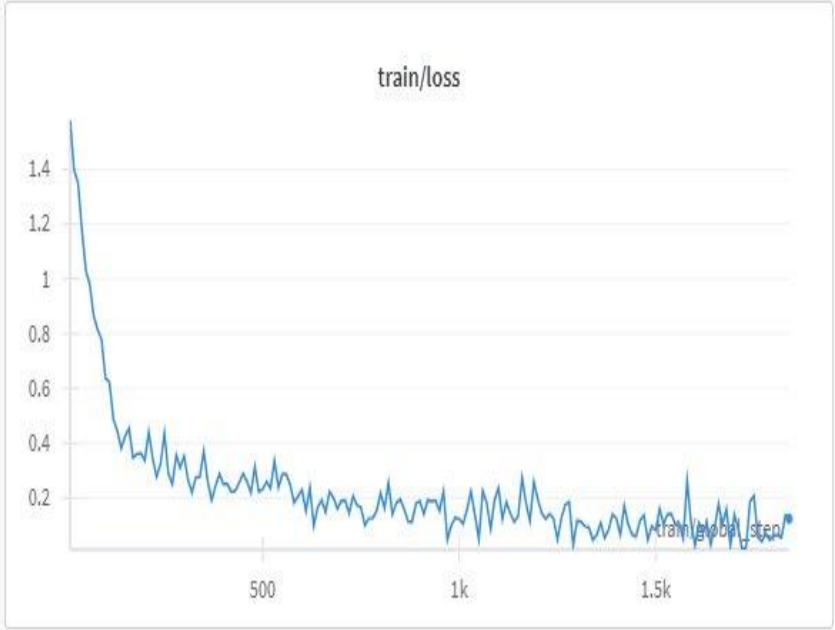
- 모델 별 평가

KLUE-RoBERTa

learning graph







프로젝트 진행과정

- 모델 별 평가

https://github.com/monologg/KoBigBird

KoBigBird

- Sparse-attention 기반의 모델로, 일반적인 BERT보다 4배 가량 더 긴 sequence를 입력할 수 있는 모델.

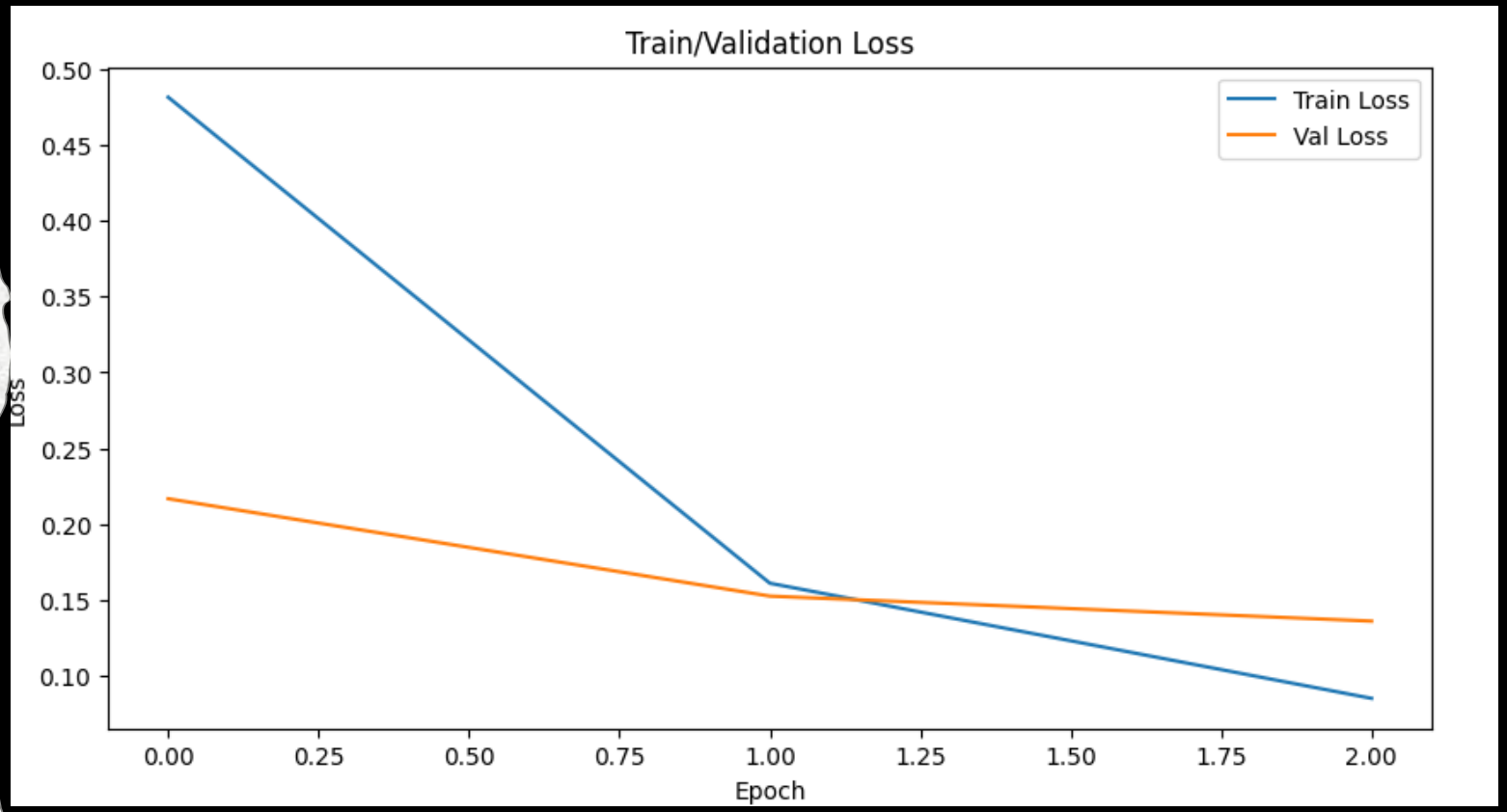
사용 모델	학습률	Optimizer
monologg/kobigbird-bert-base (하위 6개 + embedding freeze)	Lr = 2e-5	AdamW
데이터셋	**Epoch**	**결과(리더보드)**
학습데이터 (약 일반대화 1천개, 나머지 3천개)	3	0.78404

프로젝트 진행과정

- 모델 별 평가

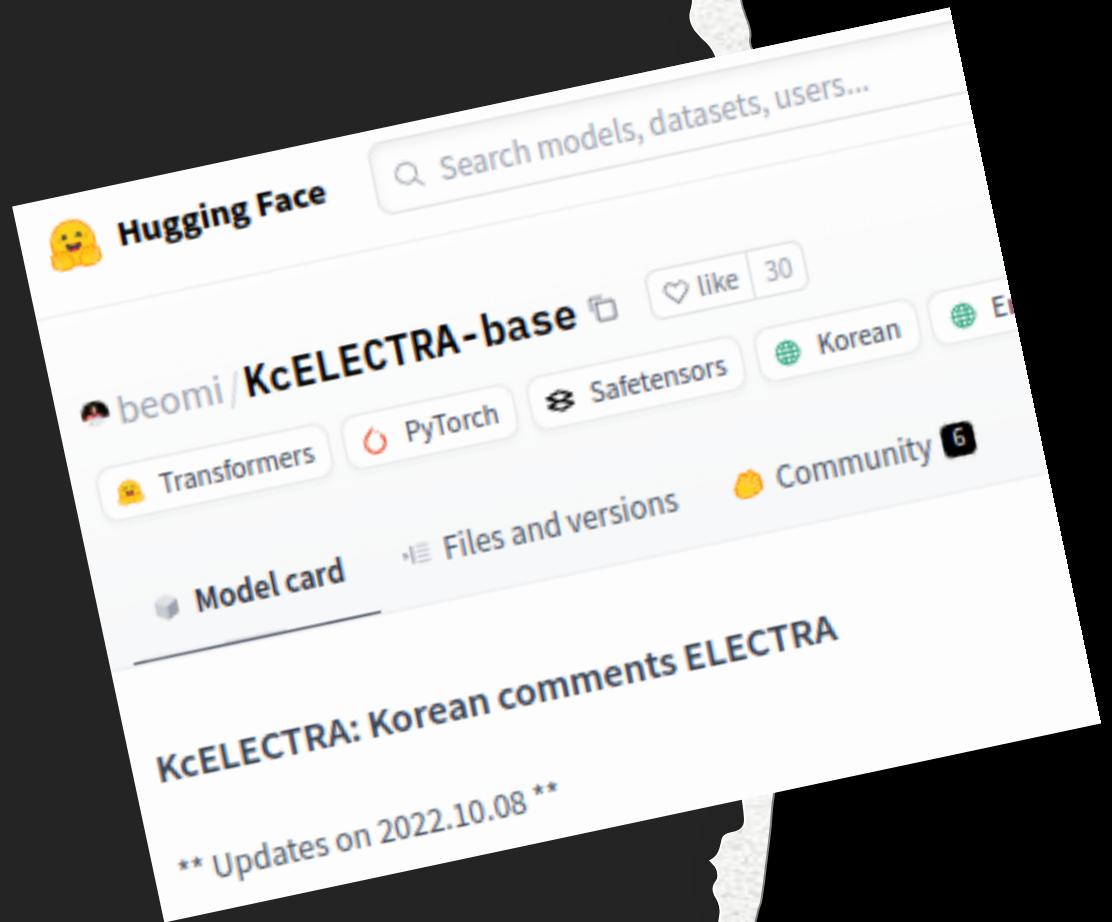
KoBigBird

learning graph



프로젝트 진행과정

- 모델 별 평가



KcELECTRA

- 기존의 BERT는 입력 토큰의 15% 가량을 마스킹해서 학습하는 Masked Language Model.
- ELECTRA는 마스킹이 아닌 대체토큰을 생성하여 모든 입력 토큰에 대해 대체 여부를 예측하는 훈련을 진행.
- BERT에 비해 경량화된 모델, 빠른 훈련속도를 보임

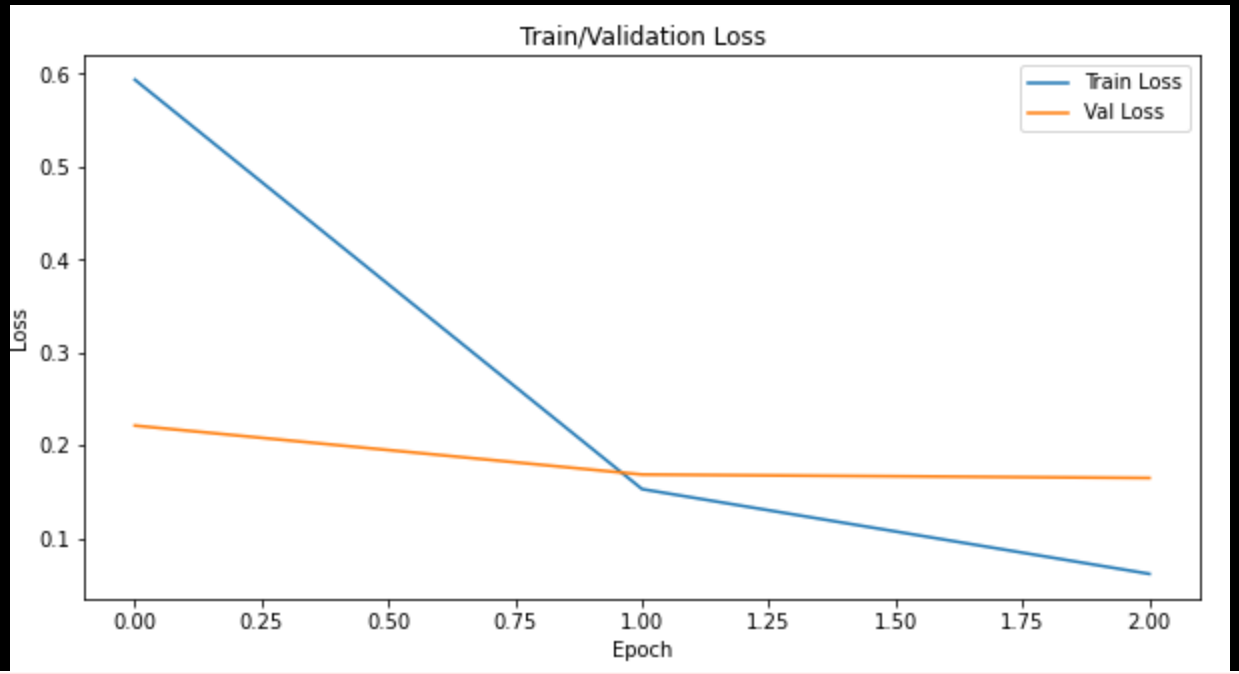
사용 모델	학습률	Optimizer
beomi/KcELECTRA-base	Lr = 2e-5	AdamW
데이터셋	**Epoch**	**결과(리더보드)**
학습데이터 (약 일반대화 4천개, 나머지 8천개)	3	0.82532

프로젝트 진행과정

- 모델 별 평가

KcELECTRA

learning graph




```
Epoch 1 [Train]: 100%|██████████| 2351/2351 [15:45<00:00,  2.49it/s]
Epoch 1 [Val]: 100%|██████████| 588/588 [01:19<00:00,  7.43it/s]
Epoch 1: Train Loss: 0.5936, Val Loss: 0.2214, Val F1: 0.9197
Epoch 2 [Train]: 100%|██████████| 2351/2351 [15:44<00:00,  2.49it/s]
Epoch 2 [Val]: 100%|██████████| 588/588 [01:19<00:00,  7.44it/s]
Epoch 2: Train Loss: 0.1530, Val Loss: 0.1687, Val F1: 0.9442
Epoch 3 [Train]: 100%|██████████| 2351/2351 [15:44<00:00,  2.49it/s]
Epoch 3 [Val]: 100%|██████████| 588/588 [01:18<00:00,  7.44it/s]
Epoch 3: Train Loss: 0.0618, Val Loss: 0.1650, Val F1: 0.9460
```

프로젝트 진행과정

- 데이터셋 증강

GPT를 활용한 증강 (위협대화)

```
# 생성 함수
def generate_augmented_conversations(base_conversations, num_samples=50):
    augmented_conversations = []

    for _ in range(num_samples):
        base1 = random.choice([
        ])
        base2 = random.choice([
        ])
        base3 = random.choice([
        ])
        base4 = random.choice([
            ""
        ])

        augmented_conversation = (
            f"{base1}₩n"
            f"{base2}₩n"
            f"{base3}"
            f"{base4}"
        )

        augmented_conversations.append(augmented_conversation)
    return augmented_conversations

# 200개
augmented_data = generate_augmented_conversations(base_conversations, num_samples=200)
```

프로젝트 진행과정

- 데이터셋 증강

구글 번역을 활용한 증강 (위협대화)

```python
from deep_translator import GoogleTranslator

def back_translate_with_deep_translator(sentences, src_lang='ko', tgt_lang='en'):
    """
    Deep Translator를 사용한 Back Translation
    """
    translator = GoogleTranslator(source=src_lang, target=tgt_lang)
    augmented_sentences = []

    for sentence in sentences:
        try:
            # 한국어 → 영어 → 다시 한국어
            translated = translator.translate(sentence)
            back_translated = GoogleTranslator(source=tgt_lang, target=src_lang).translate(translated)
            augmented_sentences.append(back_translated)
        except Exception as e:
            print(f"Error during translation: {e}")
            augmented_sentences.append(sentence)  # 실패 시 원래 문장 유지

    return augmented_sentences
```

프로젝트 진행과정

- 데이터셋 증강

데이터셋 추가 (일반 대화)


```
"info": [
  {
    "id": 41272,
    "filename": "KAKAO_898_15.txt",
    "title": "KAKAO_898_15",
    "mediatype": "SNS",
    "medianame": "카카오톡",
    "category": "일상대화",
    "date": "2021-10-05",
    "size": 886,
    "annotations": {
      "subject": "타 국가 이슈",
```


일반 데이터셋 추가

- KAKAO

```
[ ]  1 base_path = "/content/drive/MyDrive/aiffel/Data/DLthon/TS_01. KAKAO(1)"
     2 data = []
     3
     4 # i는 4(회사 아르바이트), 7(가족), 14(사회이슈)를 각각 순회하며, j는 텍스트 명 : 899부터 1797까지 전부 순회
     5 for i in [4, 7, 14]:
     6     for j in range(899, 1798):
     7         file_path = os.path.join(base_path, f"KAKAO_{j}_{i:02}.txt")
     8         try:
     9             with open(file_path, 'r', encoding='utf-8') as file:
    10                 text_data = file.read()
    11                 data.append(text_data)
    12         except FileNotFoundError:
    13             # 파일이 없으면 그냥 건너뛰고
    14             continue
```

프로젝트 진행과정

- 데이터셋 증강

1.일반 대화 데이터 증강 - 연령(청소년, 청년, 중년, 노년)별 무작위 1,000개 샘플 추출, 총 4,000개

	Unnamed: 0	연령	성별	상황키워드	신체질환	감정_대분류	감정_소분류	사람문장1	시스템문장1	사람문장2	시스템문장2	사람문장3	시스템문장3	text
304	305	중년	여성	재정, 은퇴, 노후준비	해당없음	상처	상처	지금까지 힘들게 일했는데 은퇴해서 돈이 없다고 하니 자식이 화를 내서 상처를 받았어.	돈이 없다고 하니 자식이 화를 내서 상처를 받으셨군요.	너무 화가 나. 도와줬던 건 기억 못 하고 더 받을 생각만 하니.	자식들이 받을 생각만 해서 화가 나셨군요. 이 기분을 바꿀 만한 일이 있을까요?	앞으로는 자식보다 나를 생각하며 살아야겠어. 자식 위해 살아도 필요 없다는 말 이제...	이제는 자식보다 나를 생각하며 살려 하시는군요.	지금까지 힘들게 일했는데 은퇴해서 돈이 없다고 하니 자식이 화를 내서 상처를 받았어...
305	306	중년	여성	재정, 은퇴, 노후준비	해당없음	상처	상처	친구한테 은퇴할 거라고 얘기했더니 앞으로 뭘 먹고 살 거냐면서 비웃더라고. 기분이 ...	비웃는 친구의 말에 기분이 나쁘셨군요.	아주 나쁜 친구야. 은퇴를 내가 하고 싶어 하는 것도 아닌데 말이야.	은퇴를 비웃는 친구가 나쁘다고 생각하시는군요. 앞으로 어떻게 하실 생각이신가요?	은퇴 후 조금 쉬며 즐기려고 해. 내가 잘할 수 있는 일이 분명 있을 테니까.	은퇴 후에는 쉬면서 즐기려고 할 생각이시군요.	친구한테 은퇴할 거라고 얘기했더니 앞으로 뭘 먹고 살 거냐면서 비웃더라고. 기분이 ...

2.나머지 4개 클래스에 대해서 추가적인 역변환 진행 (kor -> jap -> kor)

각 클래스별 1,094, 981, 979, 896개,총 3,950개

데이터셋 증강으로 인한 학습 데이터 수 : 기존 12300개 + 4000 + 3950 –3118(중복) = 1,7132개

프로젝트 진행과정

- 데이터 증강 후 평가

최고 성능 모델 비교 실험

사용 모델	학습률	Optimizer	데이터셋	Epoch	결과
beomi/KcELECTRA-base(하위 6개 + embedding freeze)	Lr = 2e-5	AdamW	학습데이터 (약 일반대화 4천개, 나머지 8천개)	3	0.82967
beomi/KcELECTRA-base(하위 6개 + embedding freeze)	Lr = 2e-5	AdamW	증강된 학습데이터 (약 일반대화 7.6천개, 나머지 9천개)	3	0.83252
beomi/KcELECTRA-base(하위 6개 + 상위 4 + embedding freeze)	Lr = 2e-5	AdamW	증강된 학습데이터 (약 일반대화 7.6천개, 나머지 9천개)	10	0.81297

https://wikidocs.net/165499

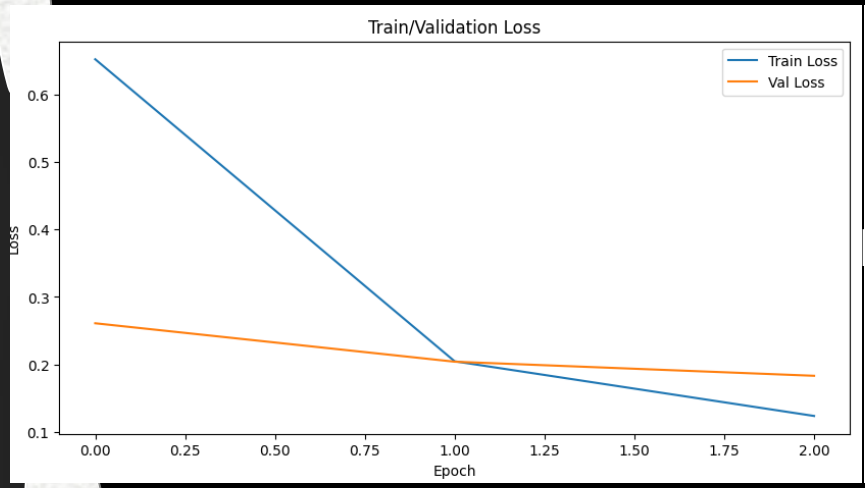
프로젝트 진행과정

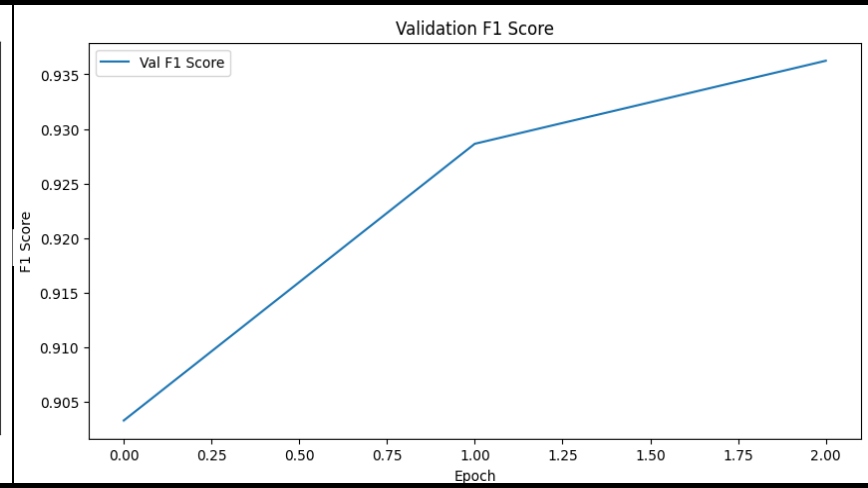
- 모델 별 평가

KcELECTRA

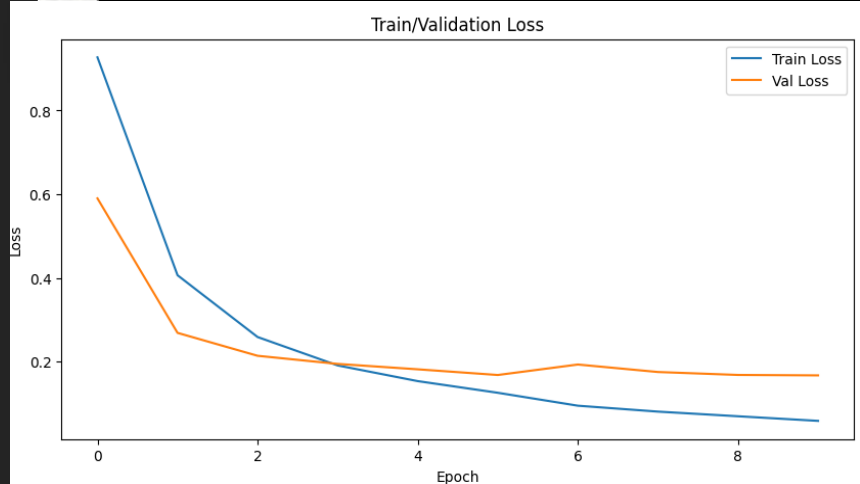
최고 성능 모델 실험 **learning graph**

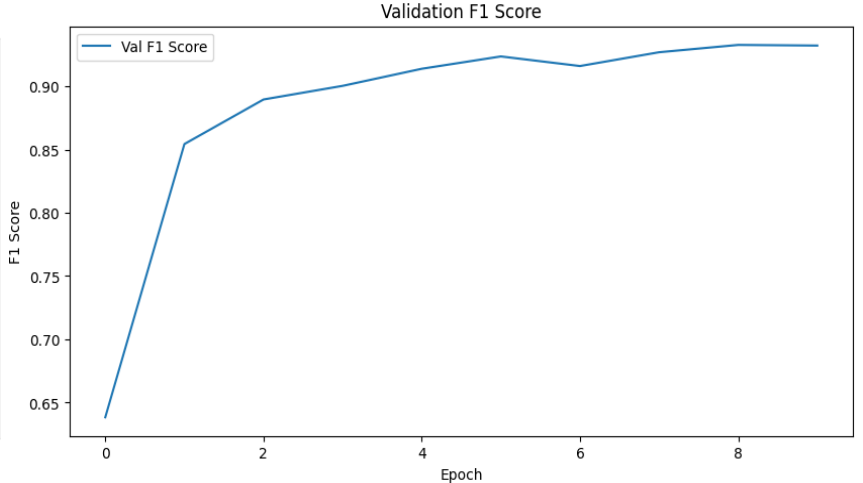
하위 6개 + embedding freeze





하위 10개 + embedding freeze





고찰

1. 추가적인 AI-Hub 데이터셋
2. 다양한 Data Augumentation 기법
    - 동의어 대체
    - 랜덤 단어 삽입, 변경, 삭제 등

1. 일반 데이터셋 맥락 파악을 위한 증강
2. 더욱 다양한 모델로 비교 분석
3. 여러 분류기를 활용한 뒤 분석
4. 학습에서 앙상블 기법 활용하여 여러 모델 섞어보기

Q & A