|  |
| --- |
| **PoC Lab – 아이펠톤 프로젝트 계획서** |

|  |  |  |  |
| --- | --- | --- | --- |
| **개발아이템명** | TuNib 기업 프로젝트 (DKTC) | | |
| **소속** | 아이펠 부산 2기 | | |
| **신청자 성명** | 진명, 심성보, 권세영 | **담당퍼실** | 김영태 |

□ 프로젝트 아이템 개요(요약)

|  |  |
| --- | --- |
| **프로젝트**  **소개** | 본 프로젝트는 TUNiB이 2021인공지능 그랜드 챌린지 4차대회 음성인지 트랙에 참가하기 위해 자체적으로 제작한 데이터셋을 이용하여 텍스트의 공격성 유형을 분류하는 프로젝트입니다. |
| **데이터의 특징** | 해당데이터는 학습 데이터와 테스트 데이터로 나뉘어 있습니다. 학습 데이터는 '협박', '갈취', '직장 내 괴롭힘', '기타 괴롭힘' 등 4개 클래스 각 1천 개 정도씩의 대화로 이루어져 있고, 테스트 데이터는 '협박', '갈취', '직장 내 괴롭힘', '기타 괴롭힘', '일반 대화' 등 5개 클래스 각 1백 개씩의 대화로 이루어져 있습니다. |
| **데이터 예시 이미지** | 이미지 |

1. 문제인식 (Problem)

**1-1프로젝트의 목표 및 목적(필요성)**

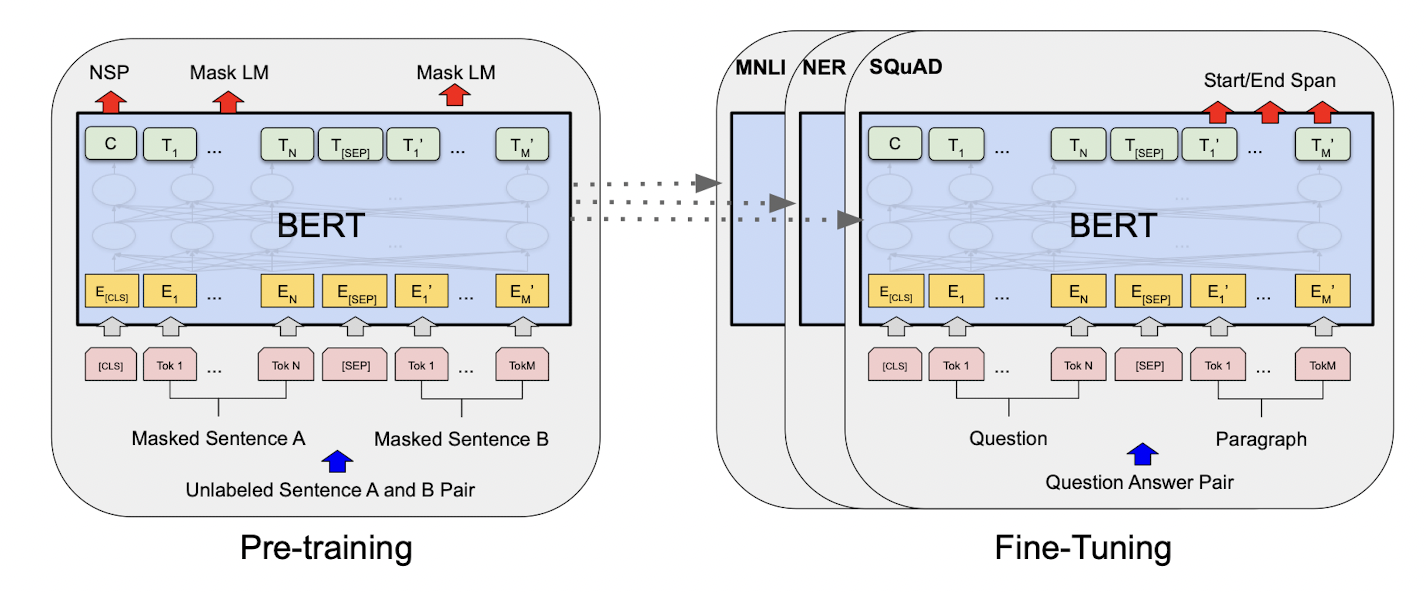
이 챌린지는 임의의 대화 음성 파일을 놓고 그 대화의 성격을 위협 세부 클래스 4개 또는 일반 대화 중 하나로 예측하는 과제입니다.

총 5개의 클래스가 들어있는 말뭉치를 입력으로 넣고 분류하는 다중분류모델을 만든뒤 F1score를 높이는 것이 해당 과제의 목적입니다.

2. 개발 및 연구 내용

**2-1. 구현 내용 상세(구현 가능성)**

1. 사용 모델



1) KLUE-RoBERTa  
 - RoBERTa 모델이 BERT를 최적화 한 모델이기때문에 성능이 좋을 것 같아서 선택함.

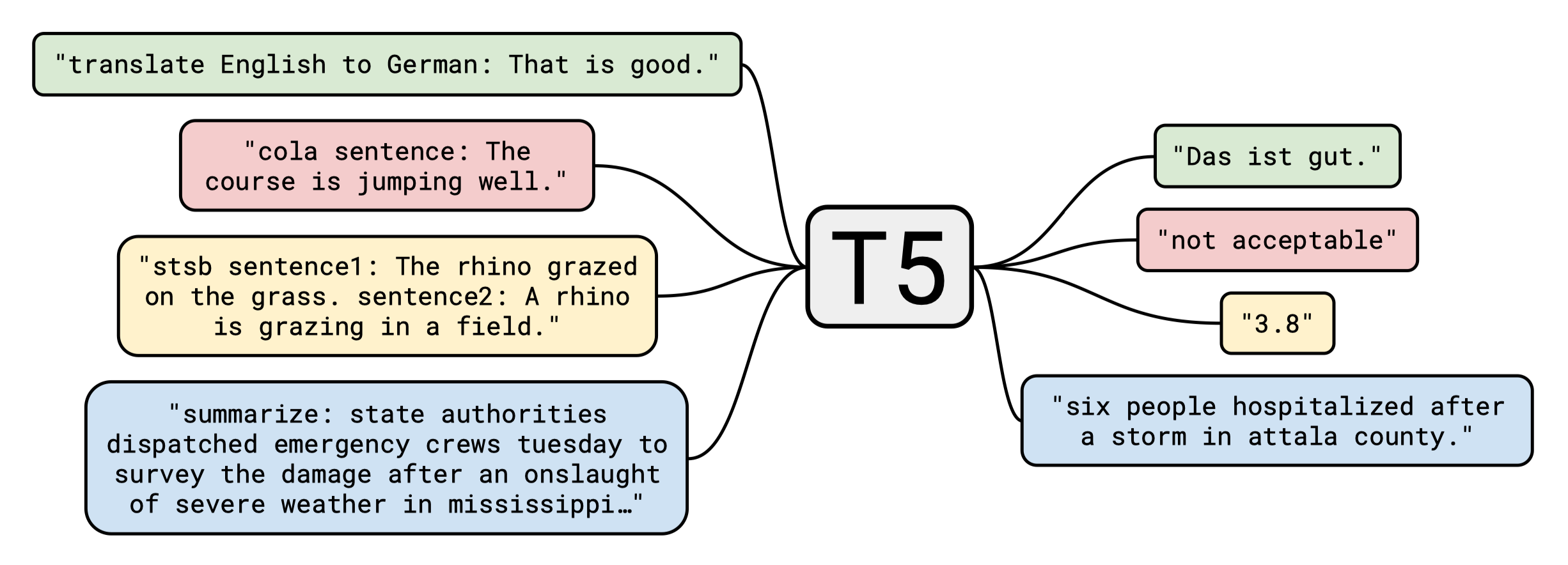
2) KLUE-BERT  
 - DKTC 설명회때 Q&A에서 KLUE BERT가 우리가 가질수있는 프로젝트환경에서 가장 퍼포먼스가 좋은 classification모델이라고해서 가져옴

3) KcELECTRA-base or large  
 - 네이버 댓글과 대댓글들을 묶은 데이터로 학습시켰기에 여러개의 댓글들이 형성하는 문맥정보를 파악할 수 있을 것 같아서 선택함.  
 - 계산량이 다른 모델에 비해 확연히 줄어들었고 좋은 성능을 보이기에 선택함.

4) KcT5

- T5가 현재 SOTA모델로 알고있어서 써보고 싶었음  
 - KcELECTRA와 마찬가지로 댓글들을 이용한 데이터셋으로 훈련시켰다는 점에서 선택함.

- 모델이 커서 사용할 수 있을지는 미지수



5) KoBERT

- 네이버 감정분석에서 좋은 성능을 보여서 선택함.

6) XLM-R-large  
 - KLUE로 사전학습된 모델중 TC성능이 가장 높은 모델

- 모델이 커서 사용할 수 있을지는 미지수

2. Augmentation

1) [nlpaug](https://github.com/makcedward/nlpaug) 라이브러리 사용. 한국어 전용은 아님.

2) 한국어 전용 라이브러리로는 [ktextaug](https://github.com/jucho2725/ktextaug)가 있음.

특히 SR(synonym replacement)과 RI(random insertion)를 사용할 때 문제가 많이 발생하는데 **제가 잘못한 건 아닌 것 같아요** 를 **제가 잘못한 총 아닌 것 같아요** (건 -> 총) 으로 바뀌기도 한다. 논문에서는 이렇게 바꿔도 꽤나 원문 데이터의 성질을 따라간다고 하지만.. 한국어의 특성상 완전히 따라가기에는 쉽지 않은 것 같다.

안전하게 데이터 증강을 하고 싶다면 RD(random deletion), RS(random swap)만을 사용하고, 데이터가 많이 필요하다싶으면 SR과 RI까지 사용하고 인간지능으로 데이터를 걸러내는 작업이 필요할 것이다.

**2-2. 개발 아이템 기대효과**

단순한 이진 분류문제가 아니라 5가지 유형에 대해 텍스트를 분류하게 되고 이러한 모델을 적용하여 친구와의 대화를 입력하여 누가 더 잘못했는가 하는 결과물을 통해 각종 테스트의 결과로 연결되는 대화 주제거리와 같이 일상 대화에 소소한 재미를 주는 서비스가 완성될 수 있을 것이다.

일단은 재미로 보는 서비스를 제공하는 것이 목적이지만, 성능이 뛰어나다면 각종 커뮤니티 사이트의 댓글관리나 분쟁 관리에도 활용할 수 있을 것으로 기대된다. (네이버의 클린봇과 비슷한 개념으로)

해당 프로젝트에서 개발된 모델을 기반으로 온라인 게임에서의 행동패턴이나 채팅 로그등을 활용하여 데이터셋을 구축하면 고의적인 욕설 유도, 트롤 행위와 같은 행동으로 인해 피해자가 억울하게 제재를 당하는 경우를 줄일 수 있을 것으로 기대되고, 꼭 욕설이 아니더라도 신조어 등을 활용하여 상대방을 비꼬거나 하는 경우에 대한 비윤리성 판단도 AI 모델을 통해 판단할 수 있게 될 것이다.

3. 실행 계획

**3-1. 기간내 프로젝트 구현 완성을 위한 전략**

◦ 2022.11.01 ~ 2022.11.09 (mini AIFFELTHON)

- 기본적인 Baseline 수립 및 테스트

◦ 2022.11.10 ~ 2022.11.13

- 추가해야할 일반 대화 데이터 수집 및 정리

◦ 2022.11.14 ~ 2022.11.20

- 일반 대화 추가 하여 실험 및 Baseline 결과와 비교  
- 추가한 일반 대화 데이터 수(1000, 5000, 10000 등)에 따른 결과 비교

◦ 2022.11.21 ~ 2022.11.27

- 추가적인 성능 향상을 방법 적용 및 실험  
- 데이터 Augmentation  
- 하이퍼 파라미터 변경

◦ 2022.11.28 ~ 2022.12.04

- 지금까지 방법 외에 성능 향상을 위한 아이디어 제시 및 구현

◦ 2022.12.05 ~ 2022.12.13

- 최종 결과 정리 및 발표 자료 준비

- 12.13 최종 발표

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 주  내용 | | M | H1 | H2 | H3 | H4 |
| 1단계 | 샘플 데이터 전처리 |  |  |  |  |  |
| 2단계 | 일반 대화 추가 후 실험 및 결과 비교 |  |  |  |  |  |
| 모델 성능 고도화 방법 적용 |  |  |  |  |  |
| 3단계 | 새로운 아이디어 제시 및 구현 |  |  |  |  |  |
| 최종 결과 정리 및 발표 |  |  |  |  |  |

**3-4. 팀장 및 팀원의 역할 분배**

|  |  |  |  |
| --- | --- | --- | --- |
| **순번** | **주요 담당업무** | **역할 상세** | **인원** |
| 1 | 데이터 수집 및 전처리 | 일반 대화 데이터셋 수집 및 전처리 | 3 |
| 2 | 실험 및 모델 고도화 | 각자 기본적인 baseline 모델 이용하여 실험 진행  이후 Augmentation이나 추가적인 아이디어 적용하여 실험 | 3 |
| 3 | 최종 결과 정리 및 발표 자료 준비 | 최종 결과 정리 및 발표 자료 준비 | 3 |

4. Reference

1. Ktextaug : <https://github.com/jucho2725/ktextaug>
2. Nlpaug : <https://github.com/makcedward/nlpaug>
3. 각종 모델 : <https://huggingface.co/>
4. 일반 대화 데이터셋 (1) : <https://aihub.or.kr/>
5. 일반 대화 데이터셋 (2) : <https://corpus.korean.go.kr/>
6. ELECTRA : <https://brunch.co.kr/@synabreu/55>
7. KoBERT : <https://github.com/SKTBrain/KoBERT>
8. TUNiB DKTC : <https://github.com/tunib-ai/DKTC>
9. KLUE Benchmark : [https://github.com/KLUE-benchmark/KLUE#klue-plms](https://github.com/KLUE-benchmark/KLUE%252525252523klue-plms)