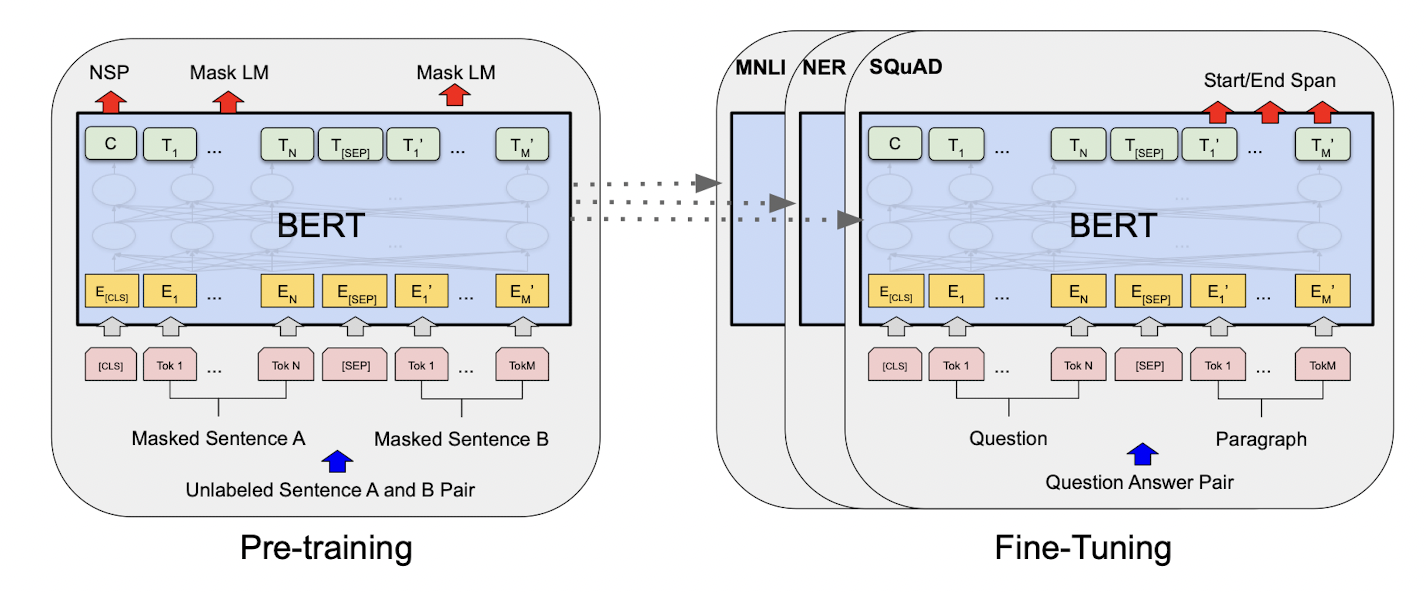
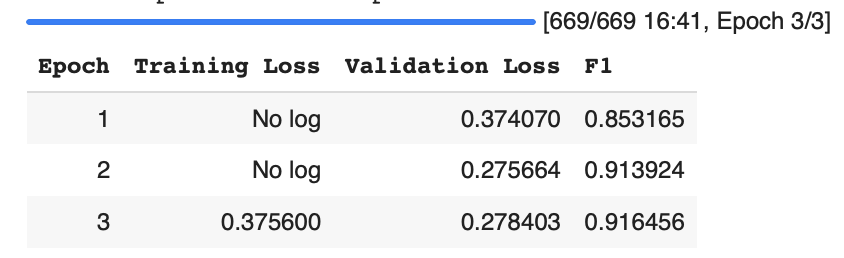
|  |
| --- |
| 『MINI AIFFELTHON 보고서』 |

|  |  |  |
| --- | --- | --- |
| **팀 명** | | 😠오빠 지금 화낸거야? |
| **제안**  **프로젝트** | **명칭** | 위험 문장 분류(DKTC) |
| **소개** | 본 프로젝트는 TUNiB이 2021인공지능 그랜드 챌린지 4차대회 음성인지 트랙에 참가하기 위해 자체적으로 제작한 데이터셋을 이용하여 텍스트의 공격성 유형을 분류하는 프로젝트입니다. |

1. **수행 및 분석**
2. *KLUE/Bert-base*

*먼저 제일 기본적인 모델로 KLUE/BERT를 사용했습니다. BERT는 GPT와 ELMO이후로 나온 denoising auto encoder모델입니다. GPT와 ELMO에서는 단 방향의 모델을 사용하여 학습했지만 BERT는 transformer의 encoder만을 사용하여 모델을 학습함으로써 양방향의 정보를 모두 활용할 수 있습니다. 현재 저희가 가지고 있는 데이터는 앞선 문장이나 토큰 이후로 어떠한 토큰이 나올지를 예측하는데 초점이 맞춰진 것이 아니라 전체적인 대화의 내용을 이해하고 그에 따라 분류를 진행해야합니다. 이에 BERT가 전체적인 문맥 정보를 단방향 정보만을 활용하는 모델보다 더 잘 파악할 수 있을거라 판단하여 사용하게 되었습니다.*

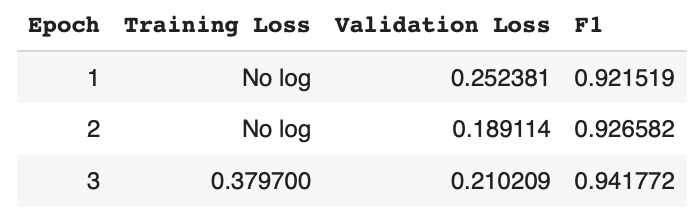
*아래 표는 주어진 학습 데이터만 이용하여 fine tuning한 결과입니다.*

**

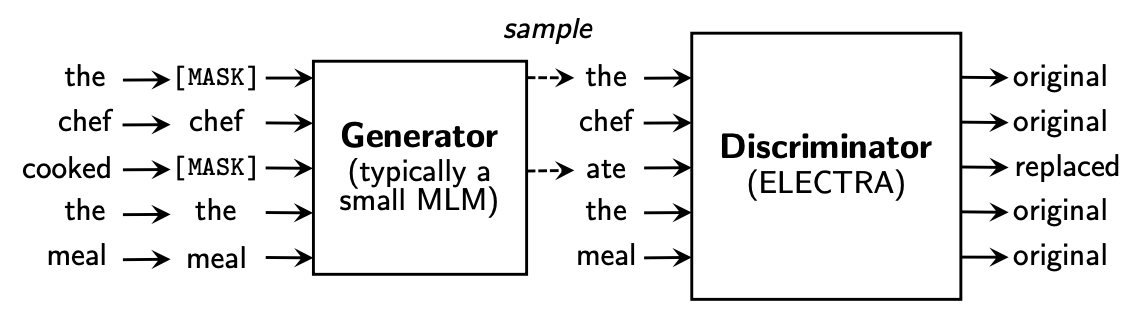
*2. KLUE/RoBERTa-base*

*다음으로 선택한 모델인 RoBERTa는 BERT가 아직 underfit되어 있다는 아이디어로 만들어진 모델입니다. 기존 BERT에 여러가지 tuning을 진행하여 기존의 BERT보다 성능을 높인 모델이기에 선정하게 되었습니다.*

*아래 표는 KLUE/RoBERTa-base로 fine-tuning한 결과입니다.*

**

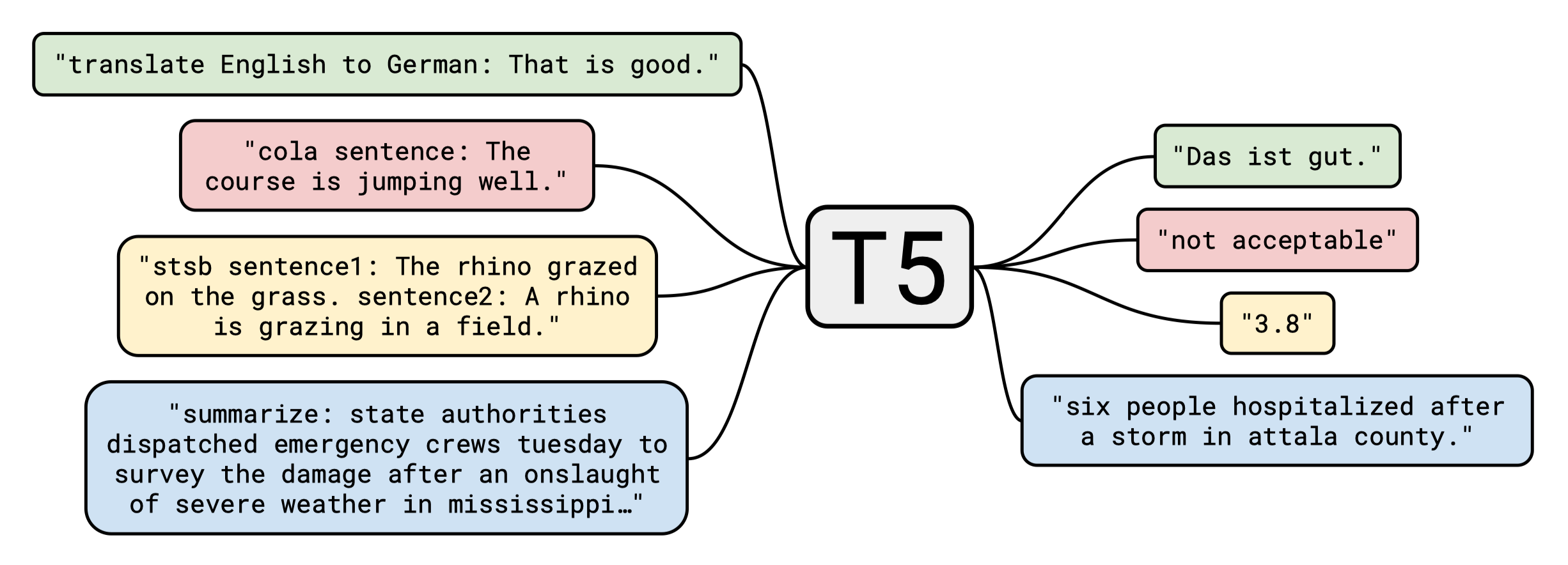
*3. KcELECTRA-base*

**

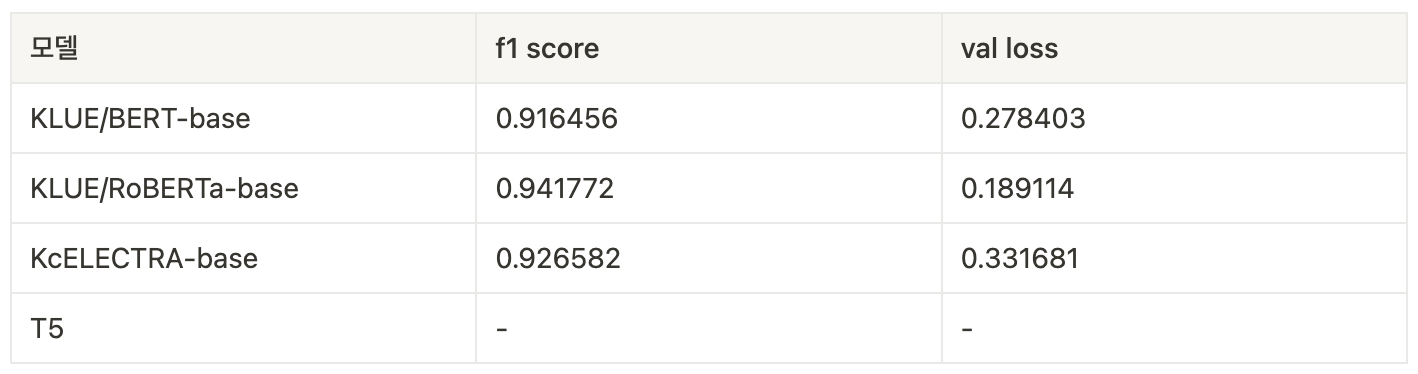
*ELECTRA는 앞선 모델과는 조금 다른 방식으로 훈련을 진행합니다. generator가 몇몇의 토큰을 다른 토큰으로 바꾸고 discriminator가 원래의 토큰인지 바뀐 토큰인지를 판별하면서 훈련을 진행하는 모델입니다. 이러한 방법으로 더 적은 계산량을 가지고 준수한 성능을 낸 모델이기에 선정하게 되었습니다.*

**

*4. T5*

*T5는 모든 NLP task를 text to text의 관점으로 수행하는 모델입니다. T5를 이용하면 하나의 모델로 다양한 task를 수행할 수 있다는 점에서 선정하게 되었습니다. 추후 실험에 사용할 아이디어로 대화내용 전체를 하나의 document로 보고 해당 대화 내용을 보다 짧은 sentence로 요약한 뒤 분류하는 아이디어가 있습니다. 이 때 따로 요약 모델을 쓰지않고 T5만을 이용하여 요약을 하고 그 후에 분류를 진행할 수 있을거라 생각하여 선정하게 되었습니다. *

*현재 진행중*

1. **결과 및 계획**
2. 결과
   1. 선택한 모델을 이용하여 일반 대화 추가 없이 결과를 한 번 뽑아봤습니다. 대체적으로 f1 score 0.9이상이 나왔습니다.
   2. T5 모델도 사용하려고 했으나 전에 사용했던 Bert, RoBerta, Electra와 다르게 encoder-decoder구조의 text2text모델이라 입력 데이터와 출력에서 약간의 차이가 있어서 간단하게 사용하기가 힘들었습니다. 현재 훈련과 label생성을 성공하기는 했지만 원하는 형식으로 나오지않아 수정이 필요한 상태입니다.

2. 미니 아이펠톤 이후 계획과 변경 사항

1. 일반 대화 추가 후 성능 확인

2. 데이터 Augmentation에 따른 성능 비교

3. 추가적인 아이디어 적용하여 성능 비교

a. Out of distribution detection

b. 대화 내용을 summarization하여 짧은 문장으로 만든 후 classification

c. 질문과 답변으로 이루어져 있는데 질문따로 답변 따로 모아서 2개의 sentence로 사 용

d. c와 비슷하지만 여기서 2개의 sentence를 사용하는 게 아니라 위협 문장에 해당하는 부분만 모아서 classification

4. 하이퍼 파라미터 조절하며 성능 비교(optuna 등 사용)

5. 위 과정을 진행하고 시간적 여유가 있으면 대화의 단순 분류가 아니라 한 대화 내에서 각 문장별로 위협적인 문장이 무엇인지를 구분하여 표시하는 task를 추가적으로 진행할 예정입니다