

DACON 한국어 문장 관계 분류 경진대회

신입기수 프로젝트

박준하 장홍선 정진호

목 차

1. 대회 개요

- 1.1 Task Definition
- 1.2 Task difficulty
- 1.3 Dataset

2. Preprocessing

3. Baseline

- 3.1 KoELECTRA : base
- 3.2 RoBERTA : base/large

4. Variations of Baseline

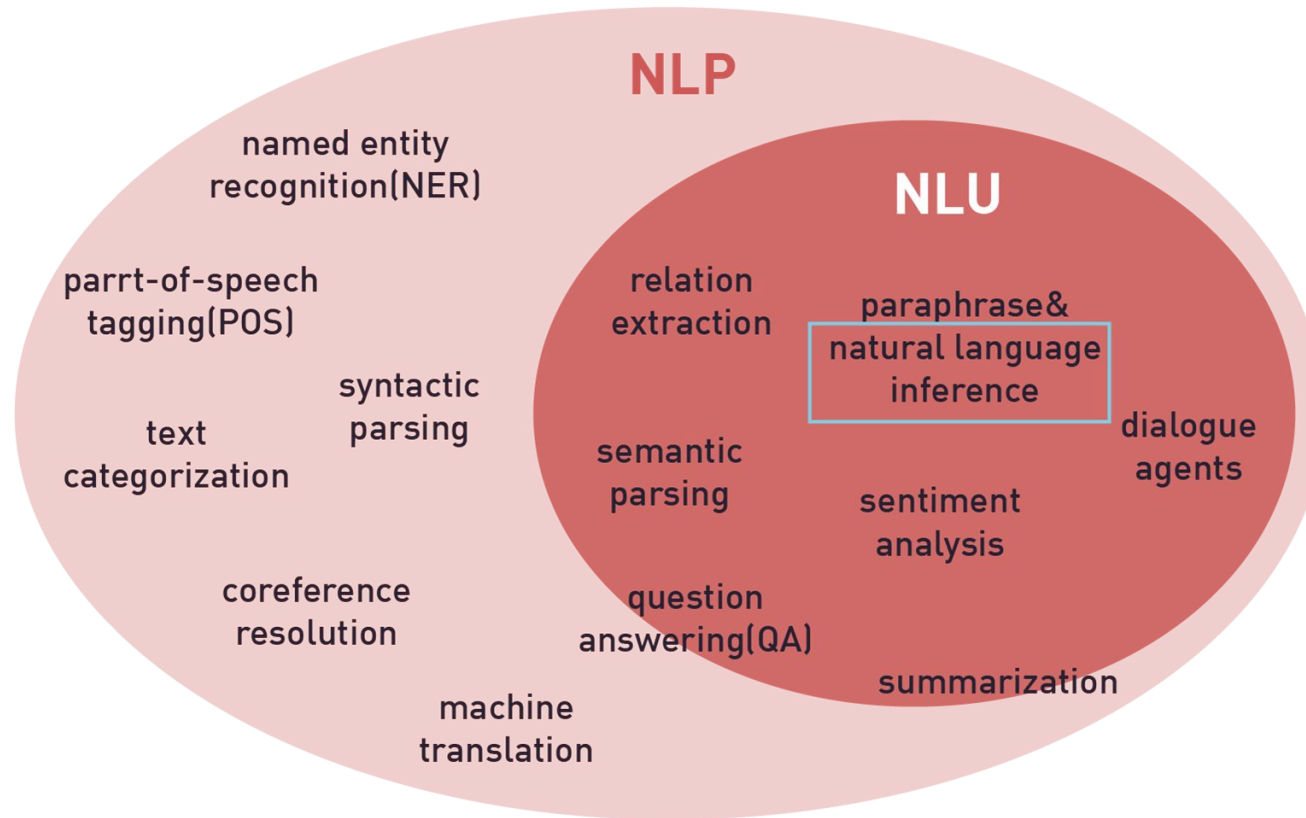
- 4.1 RoBERTa
- 4.2 ELECTRA
- 4.3 basic benchmark
- 4.4 Custom loss function
- 4.5 CLS, SEP token and maximum tokenizing length issue
- 4.6 Summary of Results

대회개요

1.1 Task Definition

1.2 Task difficulty

1.3 Dataset



- 한국어 문장 관계 분류 경진대회에서의 Task는 **NLI(Natural Language Inference)** Task이다.

Task Definition

대회 개요

- NLI Task는 Premise(전제 문장)와 Hypothesis(가설 문장)의 관계를 **Entailment(함의)**, **Neutral(중립)**, **Contradiction(기각)** 세 가지 클래스로 분류하는 Task를 의미한다. 전제 문장의 정보를 바탕으로 가설 문장의 진위를 판별할 수 있으면 그 진위 여부에 따라 Entailment/Contradiction이 되며, 진위를 판별할 수 없으면 Neutral로 분류하게 된다.

train_data.csv test_data.csv sample_submission.csv				
VIEWS Grid view Hide fields Filter Group Sort				
<input type="checkbox"/>	index	premise	hypothesis	label
1	0	씨름은 상고시대로부터 전해져 내려...	씨름의 여자들의 놀이이다.	contradiction
2	1	삼성은 자작극을 벌인 2명에게 형사...	자작극을 벌인 이는 3명이다.	contradiction
3	2	이를 위해 예측적 범죄예방 시스템...	예측적 범죄예방 시스템 구축하고 고도화...	entailment
4	3	광주광역시가 재개발 정비사업 원주...	원주민들은 종합대책에 만족했다.	neutral
5	4	진정 소비자와 직원들에게 사랑 받...	이런 상황에서 책임 있는 모습을 보여주...	neutral
6	5	이번 중설로 코오롱인더스트리는 기...	코오롱 인더스트리는 총 9만 3800톤의 생...	entailment
7	6	자신뿐만 아니라 남을 돕고자 하는 ...	모든 청년은 꿈과 열정을 가지고 있다.	neutral
8	7	시대상황을 고려하는 현명한 시청태...	시청태도에 특별한 주의점은 없다.	contradiction
9	8	사진과 차이없는 아기자기한 실내소...	아기자기한 실내소품들은 사진에서 본 것...	contradiction
10	9	빠른 답장과 간편한 체크인, 깨끗한 ...	체크인이 복잡했어요.	contradiction

- 이미 BERT, ROBERTa 등의 여러 **Pre-trained Neural Network** 모델들이 NLI Task에서 높은 성능을 보여주고 있습니다.

Model	NLI
KLUE-RoBERTa-base	84.83
KLUE-RoBERTa-large	89.17
koELECTRA-base	85.63



- 이러한 모델들이 정말로 ‘Natural Language Understanding’을 하는 것이 아니라 Premise/Hypothesis **두 문장의 subsequence matching**을 학습하고 있다는 지적 역시나 제기되었다ⁱ⁾
- 또한 NLI Task의 특성상 데이터 외적인 heuristic, 인간의 ‘상식’으로 여겨지는 정보를 활용하여야만 추론이 가능한 경우도 존재하기에, NLI Task는 여전히 몇 가지 **challenge**에 놓여 있다.

- 이어지는 문장의 경우

Sentence A : The man went to the store.

Sentence B : He bought a gallon of milk.

Label = IsNextSentence

- 이어지는 문장이 아닌 경우 경우

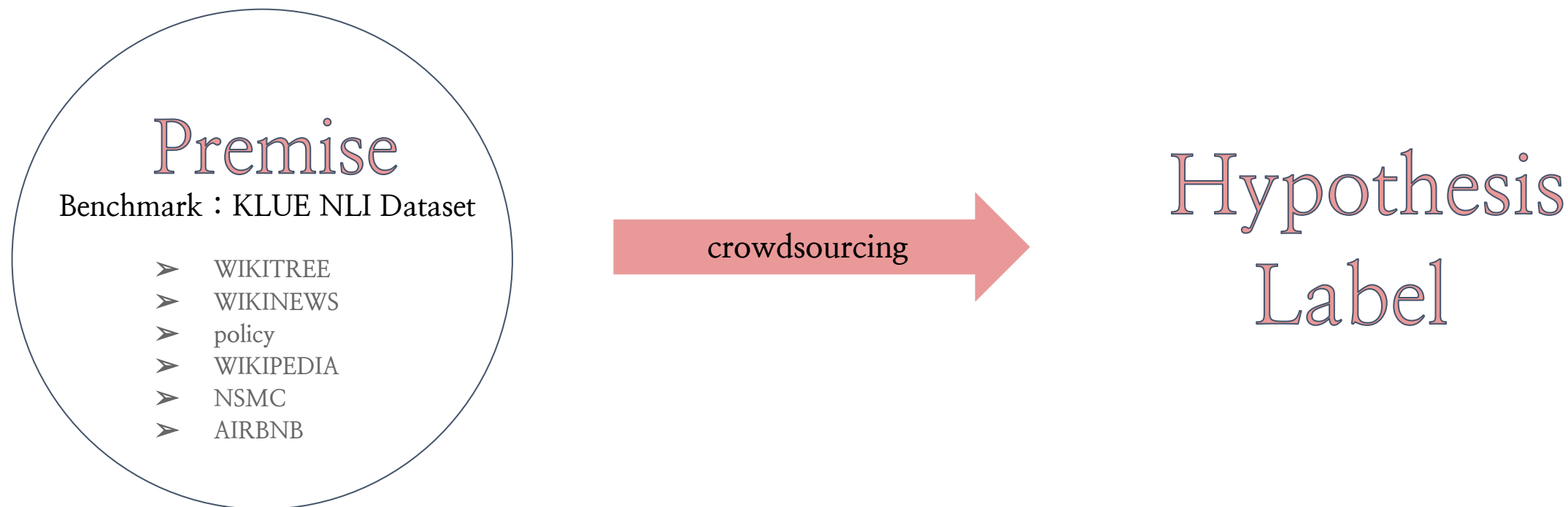
Sentence A : The man went to the store.

Sentence B : dogs are so cute.

Label = NotNextSentence



i) McCoy, R. Thomas, and Tal Linzen. "Non-entailed subsequences as a challenge for natural language inference." *arXiv preprint arXiv:1811.12112* (2018)

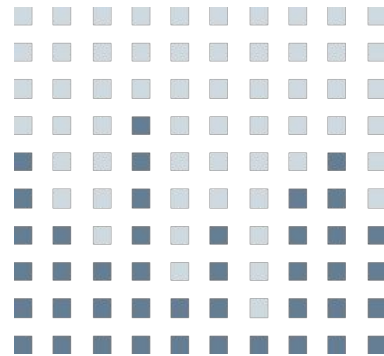


- 경진대회에서 Benchmark하고 있는 데이터셋은 KLUE NLI Dataset, 한국어로 구성된 NLI Dataset이다.
- WIKITREE, policy, WIKINEWS 세 종류의 뉴스 기사와 더불어 WIKIPEDIA(백과사전), NSMC(영화평) and AIRBNB(숙박평)를 포함해 6종류의 소스에서 Premise 데이터를 수집했다.
- **Gold Standard** : 크라우드소싱을 통해 Hypothesis 문장을 생성해 내고, 5명의 검수자가 Label을 붙여 voting을 통해 구축했다.

Source	Train	Dev	Test	Total
Wikitree	3838	450	450	4738
Policy	3833	450	450	4733
wikinews	3824	450	450	4724
Wikipedia	3780	450	450	4680
Nsmc	4899	600	600	6099
Airbnb	4824	600	600	6024
Overall	24998	3000	3000	30998

- 데이터 소스별로 Train/Dev/Test의 상대적인 비율은 동일, 대신 각 데이터셋에서 NSMC와 Airbnb의 비중이 크다.
- NSMC와 Airbnb와 같은 비격식체의 문장으로 주로 구성되어있다.
- 데이콘에서는 Dev dataset을 제외한 Train과 Test 데이터셋만을 각각 학습, 추론용으로 제공하였다.
- 본 대회에서는 KLUE Benchmark - Dev Dataset과 pretrain 모델의 사용 또한 허가되었다.

Preprocessing



“Back translation”

Data augmentation의 대표적인 방법

기존 데이터 셋을 특정 언어(영어, 일본어, 중국어)를 타겟으로 잡고
총 두 번의 번역을 통해 학습 데이터 셋을 증강 시키는 방법



papago

1. Papago API 활용

Papago 번역	Papago 번역 인공지능경망 기반 기계 번역	10,000글자/일
-----------	---------------------------	------------

- Papago API 활용 번역 시 한계점이 있다.



2. Chrome web driver 활용

- 실제 chrome browser를 통해 data를 얻기 때문에 일일 제한량 상관없이 데이터를 증강시킬 수 있다
- BUT, 1회 번역 시에 평균적으로 4초의 시간이 걸리기 때문에 주어진 데이터 1만개의 data augmentation



$$4s * 2(\text{hypothesis} + \text{Premise}) * 2(2\text{회 번역}) * 10000 = 44\text{시간}$$

→ 대회 시간 관계 상 모든 data에 대한 증강은 불가능하기 때문에 10000개의 dataset에 대해서 증강을 진행했다.



Baseline

RoBERTA : base/large
KoELECTRA : base

#	Team	Model	Description	ACC  
1		LostCow-NLI-Roberta	More	90.8
2	KLUE-team	KLUE-RoBERTa-large	More	89.17
3	KLUE-team	KLUE-RoBERTa-base	More	84.83
4	KLUE-team	KLUE-BERT-base	More	81.63
5	KLUE-team	KLUE-RoBERTa-small	More	79.33

Model	NLI
KLUE-RoBERTa-base	84.83
KLUE-RoBERTa-large	89.17
koELECTRA-base	85.63

- Baseline score는 **RoBERTa-large** → **koELECTRA-base** → **RoBERTa-base** 순이다.
- 1위의 LostCow-Roberta 모델은 Roberta-large 모델에서 약간의 하이퍼파라미터 튜닝을 거친 모델이다.
- → 따라서 **KoELECTRA-base**와 **RoBERTA-base/large**를 **베이스라인 모델**로 선정
- 비교적 가벼운 모델인 KoELECTRA에서의 실험을 통해 유의미한 성능 향상이 있는 경우 RoBERTA-large에 적용하는 방식을 채택

Variations of Baseline

4.1 RoBERTa: Base/Large

4.2 KoELECTRA

4.3 Basic Benchmark

4.4 Custom loss function

4.5 CLS, SEP token and maximum tokenizing length issue

4.6 Summary of Results

RoBERTa : Base/Large

Variations of Baseline

RoBERTa-base 모델에 대한 Experiment Result

Experiment #	Public Score	Dataset	Epoch	CV	Learning Rate
1	0.735	Vanilla	6	X	1e-6
2	0.844	Benchmark	10	X	1e-6

RoBERTa-Large 모델에 대한 Experiment Result

Experiment #	Public Score	Dataset	Epoch	CV	Learning Rate
1	0.875	Benchmark	6	X	1e-5
2	0.882	Benchmark	10	X	2e-5

Model	NLI
KLUE-RoBERTa-base	84.83
KLUE-RoBERTa-large	89.17
koELECTRA-base	85.63

KoELECTRA-SequenceClassifier Experiment Results							
Experiment #	Public Score	Dataset	Epoch	CV	learning rate	batch size	token maxle
1	0.822	Vanilla	5	5	1.00E-05	128	70
2	0.836	Benchmark	5	5	1.00E-05	128	70
3	0.839	Benchmark	10	X	1.00E-05	128	70
4	0.831	Benchmark	10	X	3.00E-06	128	70
5	0.844	Benchmark	10	X	1.00E-05	128	90
6	0.826	Benchmark	10	5	5.00E-06	16	90
7	0.857	Benchmark	10	5	5.00E-06	32	70
8	0.862	Benchmark	10	5	5.00E-06	32	128

Model	NLI
KLUE-RoBERTa-base	84.83
KLUE-RoBERTa-large	89.17
koELECTRA-base	85.63

Basic Benchmark

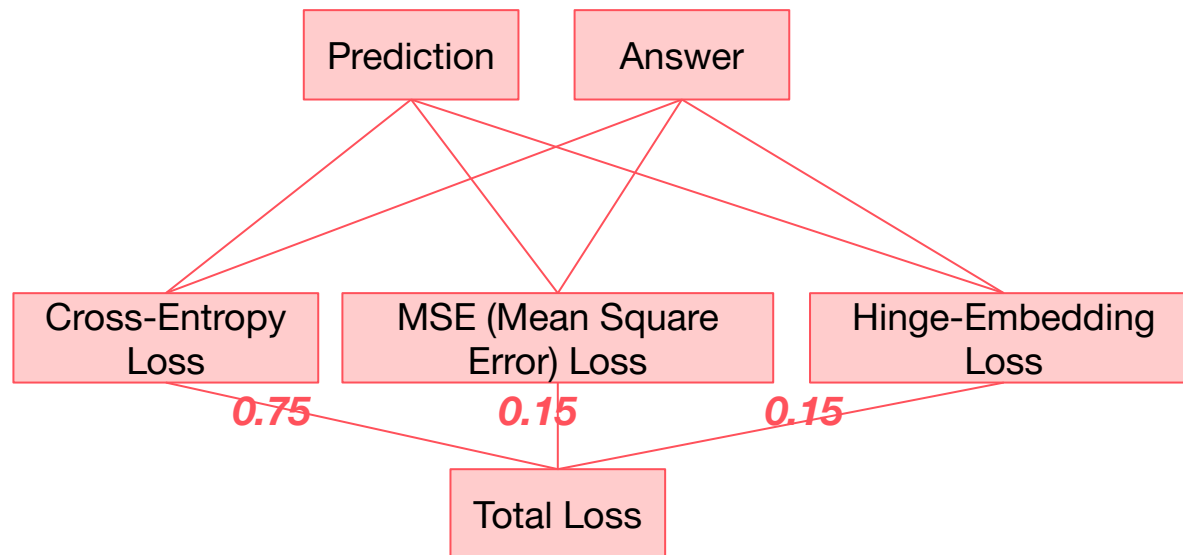
Variations of Baseline

INSIGHTS													
Dataset Effect													
1	0.822	Vanilla	5	5	1.00E-05	128	70		← Dataset Effect (+0.014)				
2	0.836	Benchmark	5	5	1.00E-05	128	70						
Epoch													
2	0.836	Benchmark	5	5	1.00E-05	128	70		← 5epoch model is underfitted (+0.003)				
3	0.839	Benchmark	10	X	1.00E-05	128	70						
Learning Rate													
3	0.839	Benchmark	10	X	1.00E-05	128	70		← Bigger learning rate is favored : local optimum				
4	0.831	Benchmark	10	X	3.00E-06	128	70						

Custom loss function

Variations of Baseline

Custom Loss Function												
3	0.839	Benchmark	10	X	1.00E-05	128	70	← Custom Loss function Enhances performance				
5	0.844	Benchmark	10	X	1.00E-05	128	70					



$\text{loss} = 0.7 * \text{loss1} + 0.15 * \text{loss2} + 0.15 * \text{loss3}$
`loss.backward()`

Hajiabadi, Hamideh, et al. "Combination of loss functions for deep text classification." (2020)

Token max length

Variations of Baseline

Batch Size vs Max Token Length												
6	0.826	Benchmark	10	5	5.00E-06	16	90	←	Batch Size effect overwhelms max token length			
7	0.857	Benchmark	10	5	5.00E-06	32	70	←				
8	0.862	Benchmark	10	5	5.00E-06	32	128		Max token length matters : CLS token problem			

- Tokenizer length has to be more than 190!
- Maximum Premise + Hypothesis length = 190

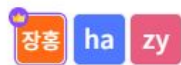
실험 결과 정리

- RoBERTa-large 모델의 경우 좋은 성능을 가지는 장점이 있지만 colab pro 환경 기준에서 학습 시 **Memory 제한으로 인한 작은 batch size**로 인해 다양한 hyperparameter에 대한 실험은 진행하지 못했다.
- 하지만 RoBERTa-large 모델을 기반으로 **KoELECTRA** 와 **robert-base** 모델에 대한 앙상블을 다양하게 **진행**했다.

Experiment	Public Score
RoBERTa-large(no CV) + KoELECTRA (CV)	0.877
RoBERTa-large(no CV) + robert-base (CV)	0.88
RoBERTa-large(no CV) + KoELECTRA (CV) + RoBERTa-base (CV)	0.883

30

장홍선



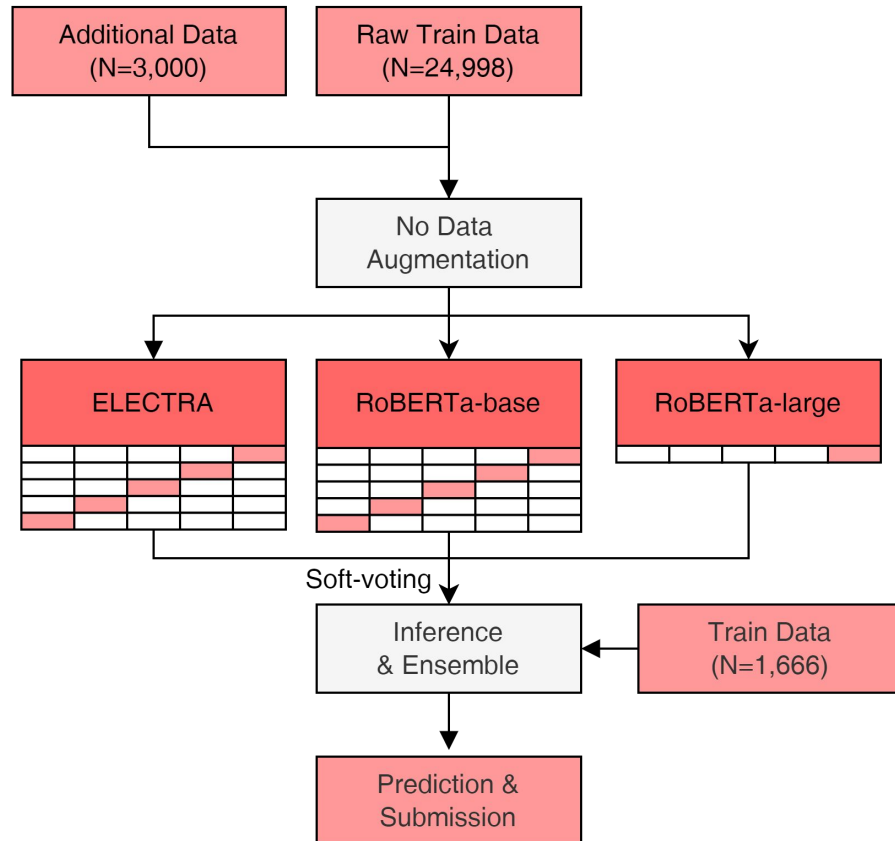
0.883

39

2시간 전

* 각자의 모델에 대해서 klue 에서 실험한 benchmark 비교에 걸맞는 성능을 획득했다.

Best(Public 0.883) Pipeline



개선 예정 사항

- 가장 좋은 성능의 roBERTa-large 모델에 대한 hyperparameter tuning + k-fold cross validation이 수행 X
- Back translation을 통한 10000개 data에 대한 augmentation 효과를 시간관계상 아직 확인 X

감사합니다