뉴스 토픽 분류
AI 경진 대회 연습 참가

백화요란 : 2022.06.07~2022.06.13

百花擾亂

목차

대회 개요

- 뉴스 토픽 분류 AI 경진 대회
- 카테고리 : 월간 데이콘 17 | 자연어 | 분류
- 기간 : 2021.06.30 ~ 2021.08.09
- 참여자 : 1069명
- 심사 기준 : Accuracy

참가 목적

- K Fold, Ensemble, Back Translation 학습 및 구현
- 기간 : 2022.06.07 ~ 2022.06.13
- Tool :
  Python / Seaborn / Translators / HuggingFace
- 참여자 :

  박정현 : 프로젝트 총괄

  권남우 : 전처리, 모델링, 보고서 및 PPT 작성

  전영욱 : 전처리, 모델링

  이성준 : 시각화, 모델링

참가 목표

대회 수상작 보다 더 나은 성능의 모델 개시

Text Classification (Accuracy)

Data Set

Back Translation

최종 3th : [Private 5위 - 0.83705 / Back Translation] (Kerry)

Stratified K Fold Ensemble (3개)

[Private 2nd] Huggingface를 사용한 베이스라인 (이강한)

Bert-base-multilingual-uncased
Klue/Roberta-base
Xlm-Roberta-large

百花擾亂

Text Classification - 데이터 셋 Introduce

Data Set

- 주제 분류를 위한 연합 뉴스 헤드라인 (YNAT)
- Klue - TC task 가공 (data shuffle / data split)
- 출처 : https://klue-benchmark.com/

Info

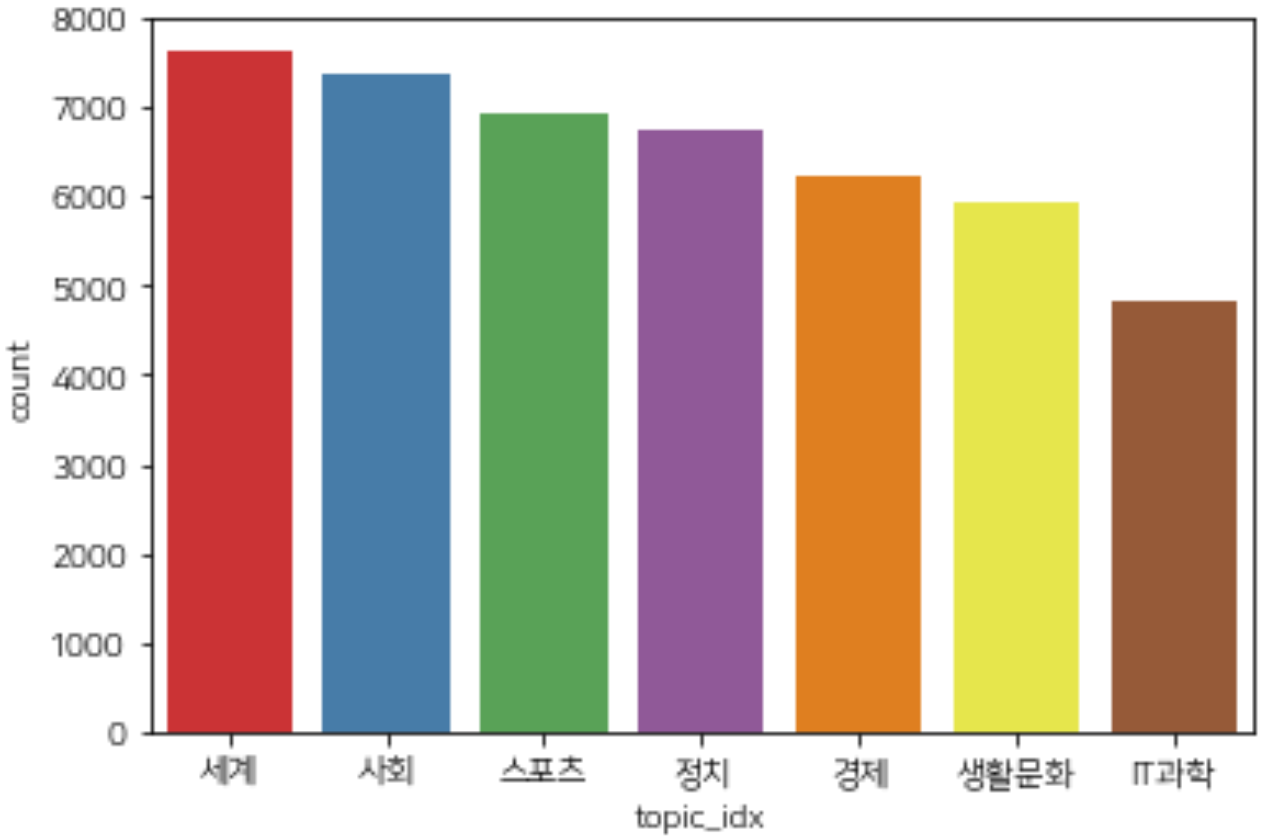
- Train : Row 45654 / Column 3
- Test : Row 9131 / Column 2
- Label
  : 0 ~ 6
  : IT과학, 경제, 사회, 생활문화, 세계, 스포츠, 정치

Column

- index : 헤드라인 인덱스
- title : 뉴스 헤드라인
- topic_idx : 뉴스 주제 인덱스 값

Text Classification - EDA

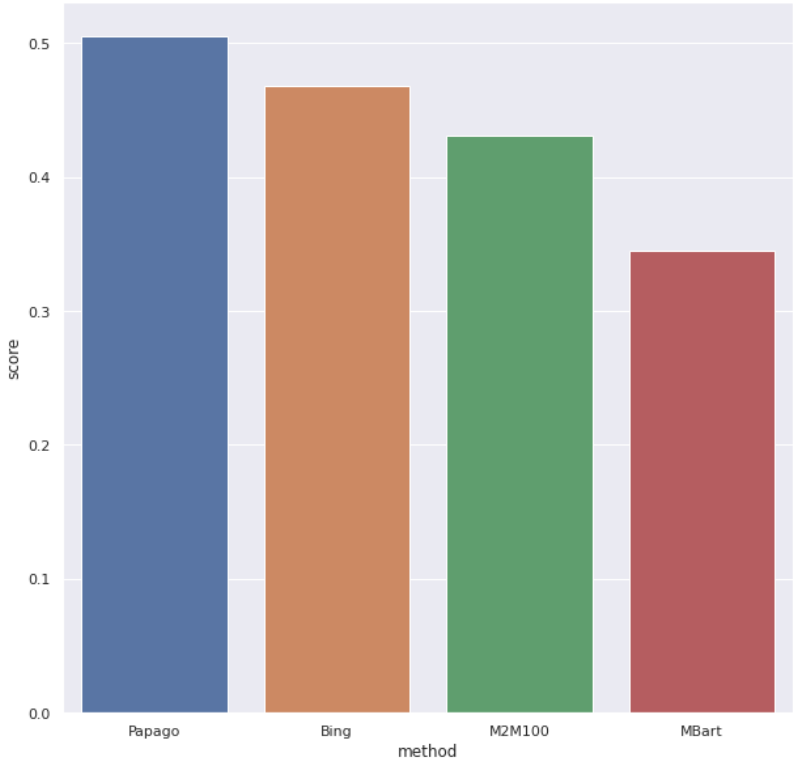
라벨 별 개수 분포



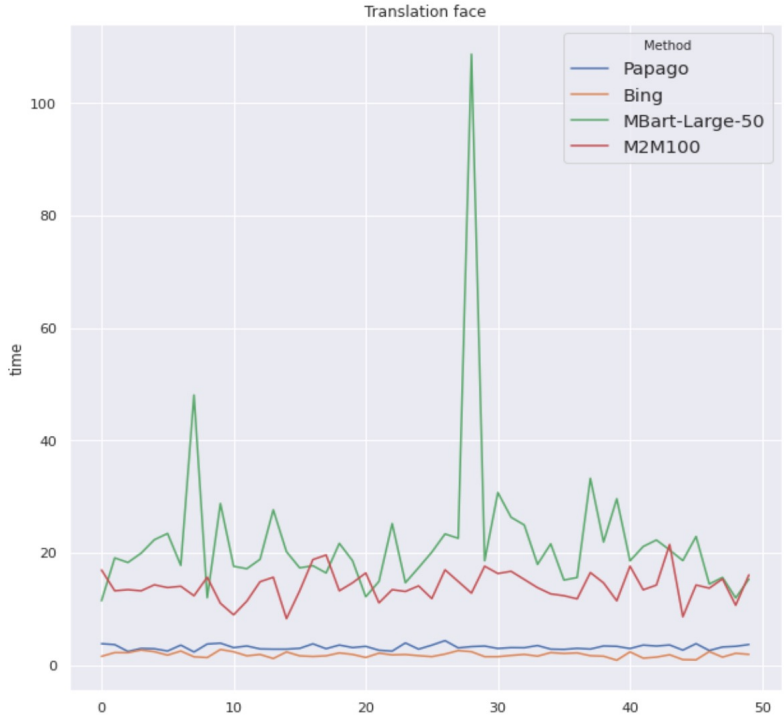
百花擾亂

Text Classification - 전처리 Back Translation (Augmentation)

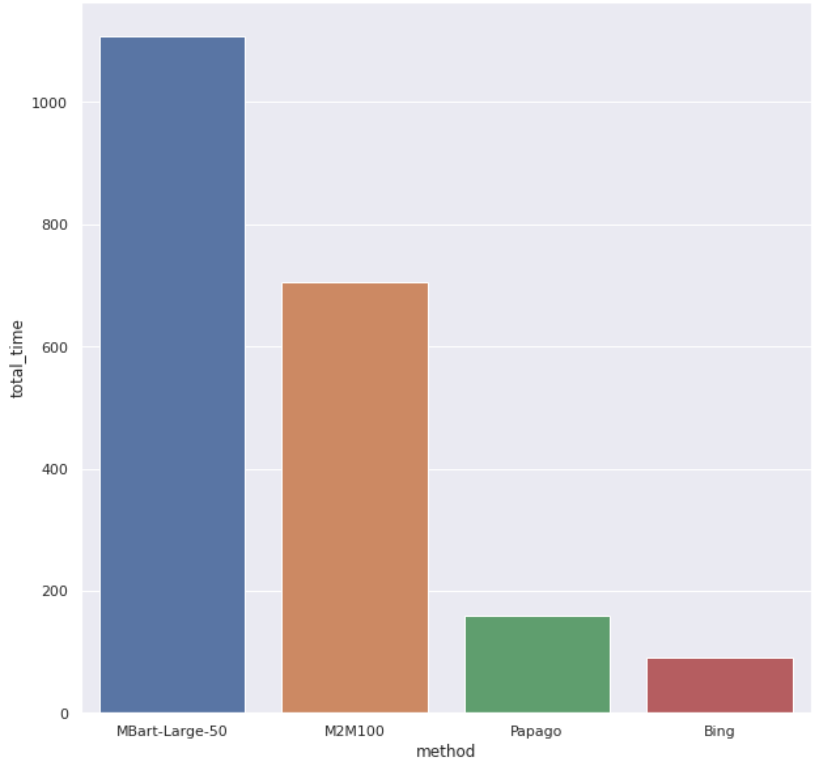
모델 별 Blue Score



문장 별 소요 시간



모델 별 총 소요 시간



* Papago 및 Bing의 경우 translators 라이브러리 활용

百花擾亂

Text Classification - 전처리 Back Translation (Augmentation)

M2M 100_418M

- 2020.10 M2M 100 모델 발표 (META)
- 세계 최초로 영어 데이터 없이 100개 언어를 번역
- 기존 영어 중심 모델보다 BLEU 메트릭스 10점 더 높음
- 총 2200개 언어 쌍 학습 / 총 154억 개 매개변수
- Hugging Face를 통해 구현 (facebook/m2m100_418M)

Process

한글 ⟶ 한글

원본 : 인천→핀란드 항공기 결항…휴가철 여행객 분통

증식 : 인천→핀란드 항공기 결말…휴가철 여행객 분열

한글 -> 영어 -> 한글 보다 빠른 속도

百花擾亂

Text Classification - 전처리 Back Translation (Augmentation)

이상치 제거

- Back Translation으로 인해 이상치 생성
- 이상치 예시 : '웃음소리 웃음소리 웃음소리 웃음소리 MVP'
- 이상치 제거 후 증식된 데이터 개수 : 35727 개

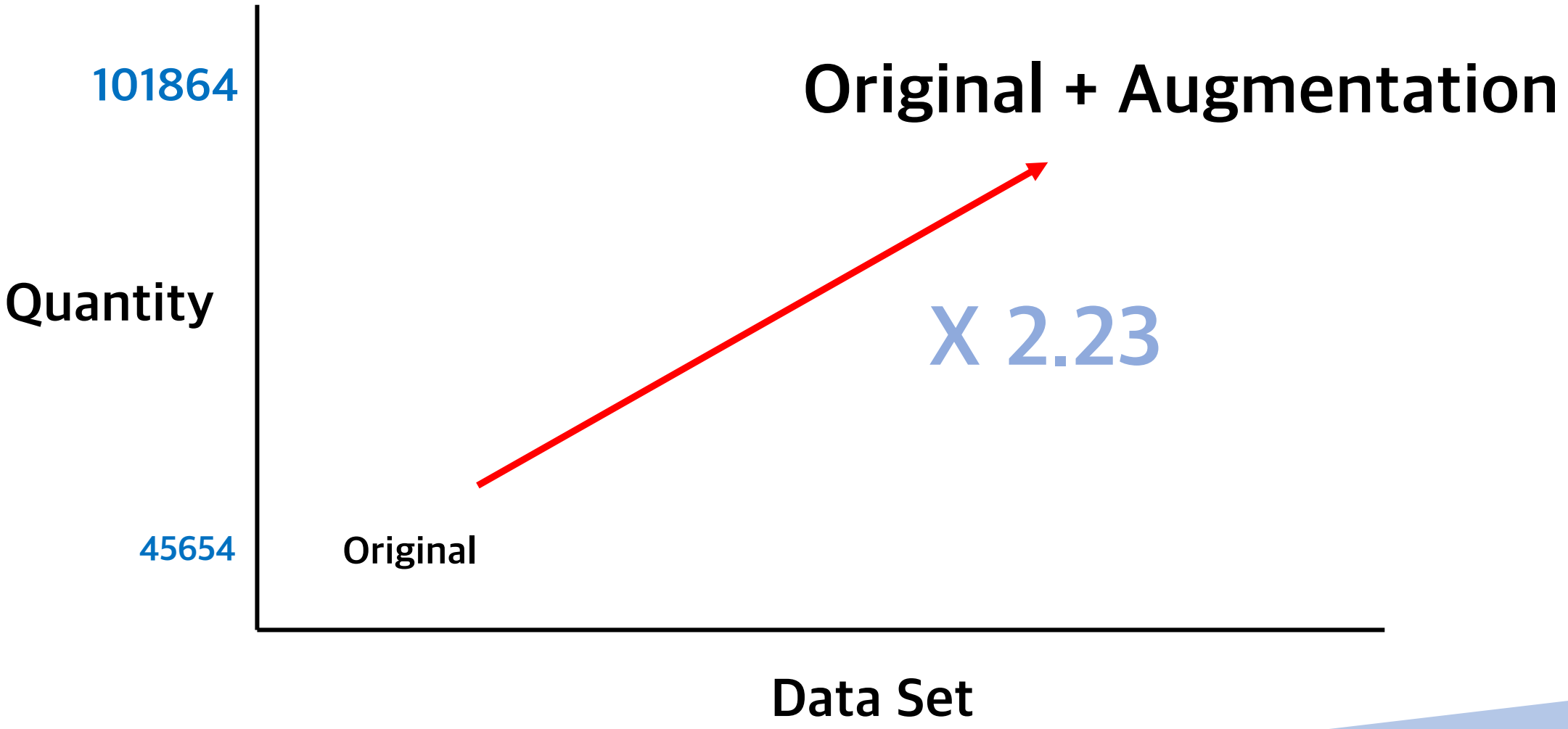
한글 -> 한자 변환

- 원본 데이터 중 한자로 표기된 단어가 존재 (EX : 美, 北, 株)
- Back Translation에 의해 모두 한글로 변환됨
- 자주 사용되는 한자를 알아내어 한자로 다시 변환

Over Sampling

- Label 별 데이터 개수가 다름을 확인
- 데이터 개수가 가장 많은 라벨 (세계 / index : 4) 기준
- 라벨 별 데이터 개수를 14552 개로 통일

Text Classification - 전처리 Back Translation (Augmentation)



Text Classification - 모델링 Stratified K Fold

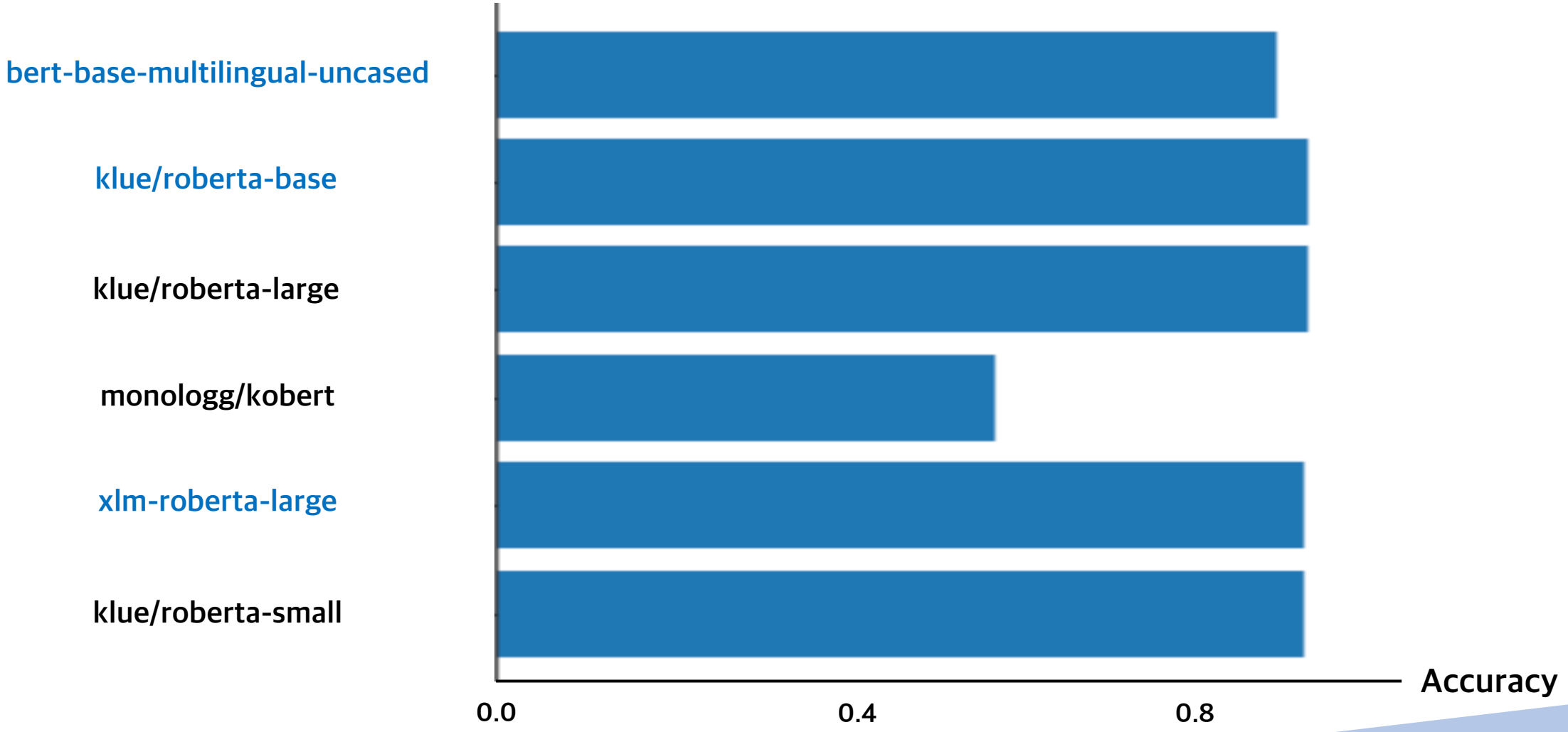
[Private 2nd] Huggingface를 사용한 베이스라인 (이강한)

- K Fold 사용 -> Label 비율이 일정하지 않은 Data Set 생성 (한계점 존재)
- K = 5로 설정
- 5 개의 데이터 셋을 6 개의 모델에 각각 학습

백화요란

- Stratified K Fold 사용 -> K Fold의 한계점 극복 (균일한 Label 비율 형성)
- K = 5로 설정
- 5 개의 데이터 셋을 3 개의 모델에 각각 학습

Text Classification - 모델링 Model



Text Classification - 모델링 Hyper Parameter

Learning Rate	LR Scheduler	Weight Decay
2e-5	Linear	0.01
Warm up Ratio	**Epoch**	**Train Batch Size** **Val Batch Size**
0.1	3	32

百花擾亂

Text Classification - 모델링 Performance

bert-base-multilingual-uncased

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.727800	0.463028	0.845482	0.845248
2	0.388900	0.402806	0.869239	0.867129
3	0.266600	0.366486	0.885829	0.885070

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.728600	0.468073	0.845384	0.843701
2	0.388700	0.376412	0.874393	0.873411
3	0.268200	0.357202	0.889314	0.888528

klue/roberta-base

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.578500	0.372346	0.875565	0.874989
2	0.289100	0.321272	0.895936	0.895691
3	0.182100	0.319363	0.905998	0.905492

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.570700	0.373928	0.875374	0.874701
2	0.294800	0.334434	0.893879	0.892574
3	0.184300	0.315677	0.906887	0.906276

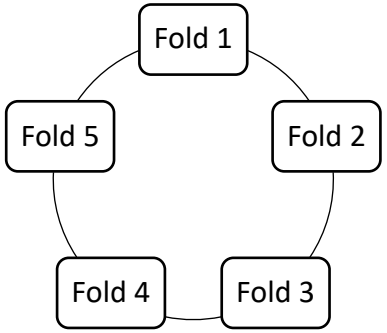
xlm-roberta-large

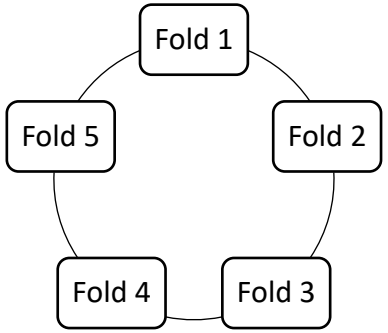
Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.690100	0.420977	0.859422	0.858226
2	0.381100	0.375175	0.876209	0.875005
3	0.279500	0.345556	0.888480	0.887787

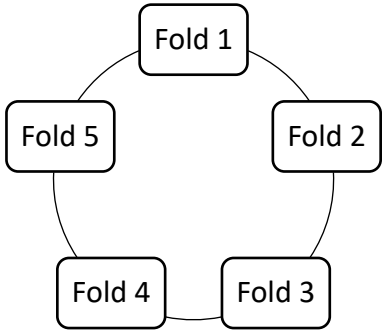
Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.694100	0.420409	0.865067	0.864196
2	0.380200	0.369074	0.880774	0.878940
3	0.280800	0.343541	0.892603	0.891848

百花擾亂

Text Classification - 모델링 Ensemble







bert-base-multilingual-uncased

klue/roberta-base

xlm-roberta-large

Logits

각각의 모델이 구한 Logits에 대한 가중치를 각기 달리함

bert-base-multilingual-uncased	klue/roberta-base	xlm-roberta-large	Accuracy
0.15	0.6	0.25	0.86878
0.15	0.55	0.3	0.86856
0.02	0.49	0.49	0.86791
0.05	0.7	0.25	0.86746

Public 20등

(256명 中)

Q & A

감사합니다