NLP SESSION

DACON 뉴스 토픽 분류 AI 경진대회

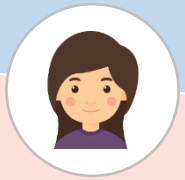
https://bit.ly/3M8Jzh9 Go

김진수

15기

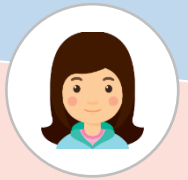
원윤정

12기



이연정

14기



제갈예빈

14기

대회소개

TASK

연합 뉴스 헤드라인을 데이터 세트로 활용해 뉴스의 주제를 분류하는 Text Classification

	index	title	topic_idx
0	0	인천→핀란드 항공기 결항…휴가철 여행객 분통	4
1	1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4
2	2	이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것	4
3	3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	4
4	4	시진핑 트럼프에 중미 무역협상 조속 타결 희망	4
5	5	팔레스타인 가자지구서 16세 소년 이스라엘군 총격에 사망	4
6	6	인도 48년 만에 파키스탄 공습…테러 캠프 폭격종합2보	4
7	7	美대선 TV토론 음담패설 만회실패 트럼프…사과 대신 빌클린턴 공격해 역효과	4
8	8	푸틴 한반도 상황 진전 위한 방안 김정은 위원장과 논의	4
9	9	특검 면죄부 받은 트럼프 스캔들 보도 언론 맹공…국민의 적	4
10	10	日 오키나와서 열린 강제징용 노동자 추도식	4

DATA

Index, title, topic_idx의 세 개의 컬럼으로 이루어짐.
train.shape : (45654,3)
test.shape : (9131,2)

Contents

EDA

- ✔ target값 분포
- ✔ 한자어 데이터 확인

Preprocessing

- ✔ 한자어 데이터 번역
- ✔ Back Translation

Modeling

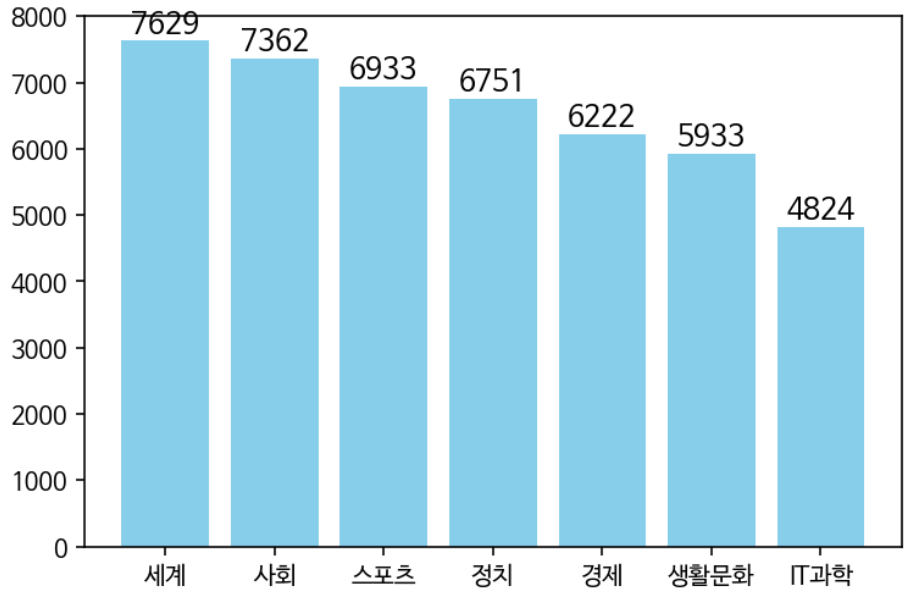
- ✔ RoBERTa 모델
- ✔ Finetuning
- ✔ Ensemble

Result

- ✔ 분석 결과
- ✔ 분석 의의
- ✔ 추후 과제

EDA

Target값 분포



카테고리 간의 분포가 전체적으로 고르진 않지만, test data의 Label 분포를 알 수 없어 Stratified K-Fold를 활용하여 Training

Preprocessing

한자어 데이터

	title	title_chinese
1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	실리콘밸리 넘어서겠다…구글 15조원 들여 미국전역 거점화
3	NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려종합	NYT 클린턴 측근한국기업 특수관계 조명…공과 사 맞물려종합
7	美대선 TV토론 음담패설 만회실패 트럼프…사과 대신 빌클린턴 공격해 역효과	미국대선 TV토론 음담패설 만회실패 트럼프…사과 대신 빌클린턴 공격해 역효과
10	日 오키나와서 열린 강제징용 노동자 추도식	일본 오키나와서 열린 강제징용 노동자 추도식
13	美올랜도 병원 최악 총기 테러 부상자 치료비 안 받는다	미국올랜도 병원 최악 총기 테러 부상자 치료비 안 받는다
...	...	...
45603	北 고난의 행군 이어 군자리정신까지…대북제재 효과	북한 고난의 행군 이어 군자리정신까지…대북제재 효과
45611	北 연이은 도발에 靑 신속·단호 대응…내부선 수위조절 고민도	북한 연이은 도발에 청와대 신속·단호 대응…내부선 수위조절 고민도
45618	실사판 옥자 나오나…中 돈육대란에 초대형 돼지 사육 봄	실사판 옥자 나오나…중국 돈육대란에 초대형 돼지 사육 봄
45623	선거 유세 나선 日 입헌민주당 에다노 대표	선거 유세 나선 일본 입헌민주당 에다노 대표
45638	日자민당 원로 헌법9조는 세계유산…개정 바늘귀만큼도 안돼	일본자민당 원로 헌법9조는 세계유산…개정 바늘귀만큼도 안돼

한자어가 존재하는 데이터의 수 : 6303개

뉴스 제목의 특성상 한자를 포함한 데이터 多
➡등장 빈도가 10 이상인 한자어에 대해
번역 실행

Preprocessing

Back Translation

Back Translation 이란?

한글로 이루어진 데이터를 영어로 번역한 후
이를 다시 한글로 번역하여 컴퓨터가 이해하기
쉽도록 문장을 변형하는 방식

Back Translation의 효과

① 뉴스 제목의 특성 상 완결된 어미로 끝나지
않는 경우 多 ➔ 완결된 문장으로 번역 가능
② Data Augmentation

```python
# Crawling
def kor_to_trans(text_data, trans_lang,start_index,final_index):

  target_present = EC.presence_of_element_located((By.XPATH, '//*[@id="txtTarget"]'))

  for i in tqdm(range(start_index,final_index)):
    if (i!=0)&(i%99==0):
      time.sleep(2)
      #print('{}th : '.format(i), backtrans)
      np.save(data_path+'kor_to_eng_train_{}_{}.npy'.format(start_index,final_index),trans_list)

    try:
      driver.get('https://papago.naver.com/?sk=ko&tk='+trans_lang+'&st='+str(text_data.iloc[i]))
      time.sleep(1.5)
      element=WebDriverWait(driver, 10).until(target_present)
      time.sleep(0.1)
      backtrans = element.text
      print('{}th : '.format(i), backtrans)
      print(str(text_data.iloc[i]))

      if (backtrans=='')|(backtrans==' '):
        element=WebDriverWait(driver, 10).until(target_present)
        backtrans = element.text
        trans_list.append(backtrans)
      else:
        trans_list.append(backtrans)

    except:
      trans_list.append('')
```

Selenium을 활용하여 PAPAGO 번역기 크롤링

Preprocessing

Back Translation

```
1 trans_list=[]
2 eng_to_trans(need_more['eng_title'], 'ko',0,len(need_more))
3 np.save('/content/drive/MyDrive/KUBIG/WINTER/KUBIG CONTEST/data/eng_to_kor_not_yet.npy',trans_list)
```

```
100% 8/8 [00:17<00:00, 2.04s/it]
0th :  품질을 떨어뜨리는 숙박음식 대여...비은행 비중이 30%에 육박해 가장 높다.
Lending accommodation food that deteriorates quality...The proportion of non-banks is close to 30percent, the highest.
1th :  대학 졸업 예정자의 24.8%는 졸업 전에 취업한다.그것은 작년보다 3.8퍼센트 증가했다.
24.8percent of prospective college graduates are employed before graduation...It increased by 3.8 percent from last year.
2th :  정경 코바체비치 25년 만의 공연...100% 음악만.
Jeonggyeong Kovachevich's first performance in 25 years...100percent music only.
3th :  페트로브라 석유 부족 딜레마...80%의 세습. 사업성 질문.
Petrobras Low Oil Dilemma...80percent hereditary. Business feasibility question.
4th :  HDC현대개발의 3분기 영업이익은 전년 동기 대비 1326억원 41.4% ↑
HDC Hyundai Development's third-quarter operating profit is KRW 132.6 billion.41.4percent compared to last year ↑
5th :  삼성 C
Samsung C&T's first-quarter operating profit is 147 billion won.It increased by 39.8percent from last year.
6th :  한국감정원이 지난 9월 전통시장 상품권으로 지급한 월급의 10%를 지급했다.
10percent of the salary paid by the Korea Appraisal Board in September as a gift certificate for traditional markets.
7th :  바른테크놀로지 스포라이브 주식 52억원 인수...8.1%의 점유율.
Acquired 5.2 billion won in shares of Barun Technology SpoLive...Share of 8.1percent.
```

번역 & 역번역 과정에서 %, &와 같은 기호가 존재할 경우 번역이 완벽하게 되지 않은 경우 有
➡ 해당 데이터 추출 후 기호를 언어로 바꿔 재번역

Back Translation 결과

title_chinese	eng_title	re_kor_title
네이버 지식인에 물어보면 인공지능이 답한다	If you ask Naver's intellectual, artificial in...	네이버 지식인에게 물어보면 인공지능이 답할 것이다.
올해 여름 평년과 비슷하거나 높은 기온...태풍은 2개 예상종합	Temperatures similar to or higher than average...	올 여름 평년과 비슷하거나 더 높은 기온...예상되는 태풍은 두 가지가 있습니다.
신간 선한 영향력	Fresh influence.	신선한 영향.
허경민 역전 3루타 두산 롯데 꺾고 3연승...곽빈 데뷔...	Heo Kyung-min's come-from-behind triple beat D...	허경민의 역전 3루타 두산 롯데에 3연승...Kwak Bin의 데뷔...
올해 최고 과학성과 중성자별 충돌 입증·신종 오랑우탄종합	This year's best scientific achievement, proof...	올해 최고의 과학적 업적, 중성자 충돌 증명, 종합 오랑우탄.

역번역된 결과가 비교적 이해하기 쉽게 풀어서 해석됨.
[한계] 최종적으로 역번역 된 데이터를 확인해보면 영어와 한국어가 혼재된 경우 有

Modeling

Model

Roberta 모델

Roberta 모델이란?

기존 텍스트 분류에서 많이 활용되던 BERT 모델을 향상시킨 모델

① 더 많은 데이터로 더 오래, 더 큰 배치로 학습
- BERT(16GB) ➡ RoBERTa(160GB)
- 학습데이터 중 뉴스 데이터 또한 포함

② “Dynamic Masking” 활용
- 매 epoch마다 다른 Masking 실행

③ Byte-level BPE 활용
- BPE(Byte-pair Encoding) 이란?
- 단어들을 모두 글자 단위로 분리한 다음 기존 단어 딕셔너리를 참고하여 단어 내 sub-word unit으로 분리하는 방식

Modeling

How?

① 학습 시 한글 데이터만 활용한 것이 아니라 1차 가공된 영어 데이터도 함께 활용하여 학습

② 팀을 나누어 영어 데이터 / 한글(기존+역번역) 데이터를 각각 학습

한글 : klue/roberta-small/base 모델 활용

```
self.bert = TFRobertaModel.from_pretrained("klue/roberta-small", from_pt=True)
self.dropout = tf.keras.layers.Dropout(self.bert.config.hidden_dropout_prob)
self.classifier = tf.keras.layers.Dense(num_class, kernel_initializer=tf.keras.initializers
                                        name="classifier")
```

영어 : distilroberta-base/roberta-small, bert-base 활용

```
self.bert = TFRobertaModel.from_pretrained("distilroberta-base", from_pt=True)
self.dropout = tf.keras.layers.Dropout(self.bert.config.hidden_dropout_prob)
self.classifier = tf.keras.layers.Dense(num_class, kernel_initializer=tf.keras.initializer
                                        name="classifier")
```

③ 학습 결과를 Ensemble 하여 최종 결과 도출

각각 학습된 모델을 제출하여 얻은 public 점수를 가중치로 활용하여 가중합 실시

```
#1 public 확인 및 softmax

pub_jin = 0.8701
pub_yoon = 0.7363
pub_ye = 0.8664
pub_yeon = 0.85

p1 = softmax([pub_jin, pub_yeon, pub_ye, pub_yeon])
```

```
1 #2 public 가중치와 각 모델의 가중치 곱하기
2 jin = jinsu*p1[0]
3 yoon = yoonjung * p1[1]
4 ye = yebin * p1[2]
5 yeon = yeonjung * p1[3]
```

```
1 import tensorflow as tf
2 result = jin + yeon + yoon + ye
3 result = tf.argmax(result,axis=1)
```

Result

최종 결과

Public :	0.87009	17위	한글(원본+역번역) 데이터 roberta-base 단독 모델 사용
Private :	0.83552	8위	한글(원본+역번역) roberta-base & roberta-small & 영어 bert-base 세가지 모델 앙상블

결과 분석

Back translation에서 최종적으로 역번역된 결과뿐만 아니라 1차 번역된 영어 데이터 역시 성능 향상에 기여

추후 과제

① 보다 더 세심한 전처리
ex. 한글과 영어가 혼재된 문장 처리…
② roberta-large와 같은 보다 더 학습된 모델 활용

감사합니다