Text Classification With Roberta

TEAM : 2팀 나는 자연인이다.

김태웅, 설재민, 이재형, 이세린

CONTENTS

1 Project concept

Introduction
-
Analysis process

2 Project Team

3 Project Process

4 Project Result

EDA
-
Model Selection
-
Model Improvement

5 자체 평가 및 보완

1
Project Concept

Project concept

Introduction

영어 문장에 대한 긍정/부정 리뷰를 구분하는 **text classification** task입니다.

사용 모델 : 다양한 모델을 프로젝트에 적용시켜 최적의 모델을 찾는다.

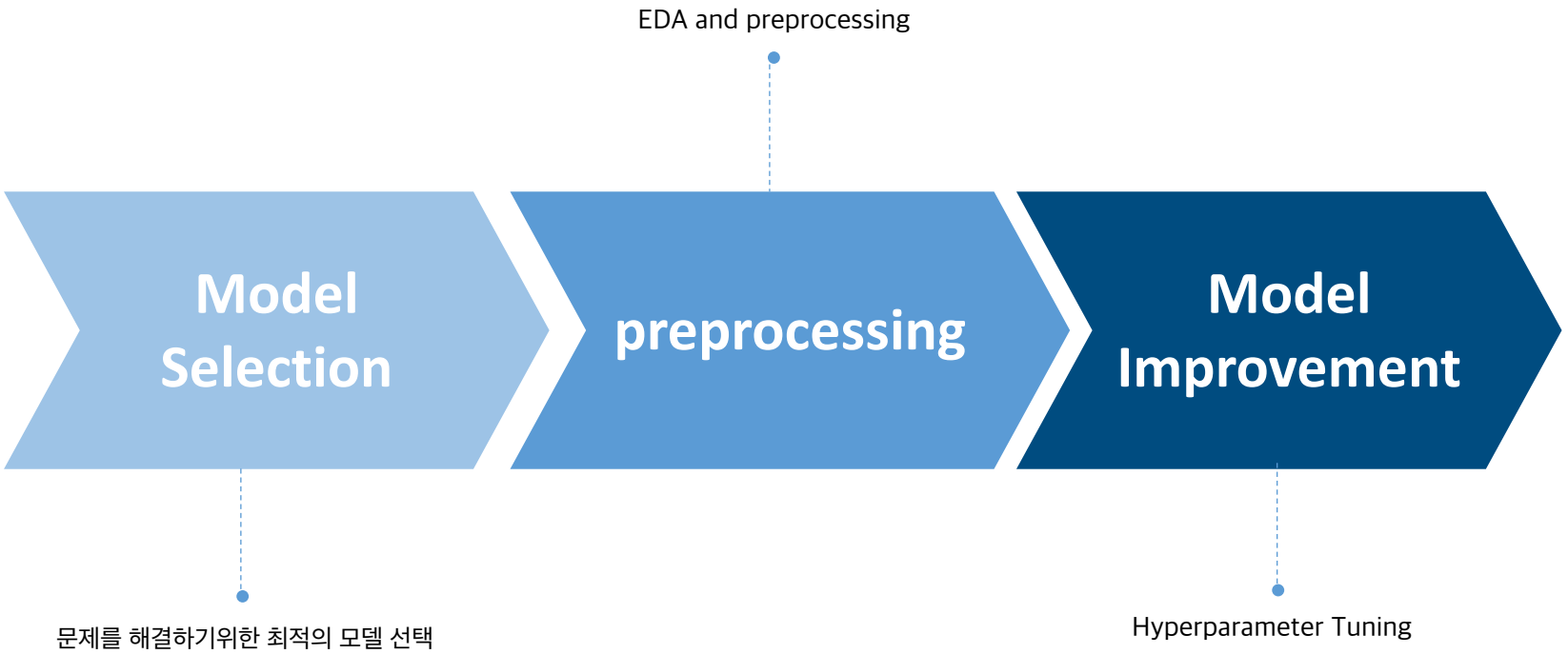
- Bert
- **RoBerta** : 테스트 결과 가장 성능이 좋게 나왔다.
- Bart
- Albert
- XLNet

Framework Google colab

Library PyTorch HUGGING FACE

Project concept

analysis process



2
Project Team

Project Team

김태웅(팀장)	Project process 설계, RoBERTa 모델 적용, 데이터 전처리 및 시각화
설재민(팀원)	bert 모델 적용 및 개선
이재형(팀원)	Bart 모델 적용 및 개선
이세린(팀원)	XLNet 모델 적용 및 개선, Roberta 모델 하이퍼 파라미터 튜닝

3
Project process

Project Process

구분	기간	활동
데이터 파악	10/25~10/26	베이스라인 코드 파악
모델 선정 및 분석	10/26~10/28	Xlnet, bert, bart, Roberta를 통한 모델링
데이터 전처리 및 시각화	10/28~10/29	EDA, preprocessing
하이퍼 파라미터 튜닝	10/29~10/31	Batch size, epoch, learning rate tuning
총 분석기간	10.25~10/31	

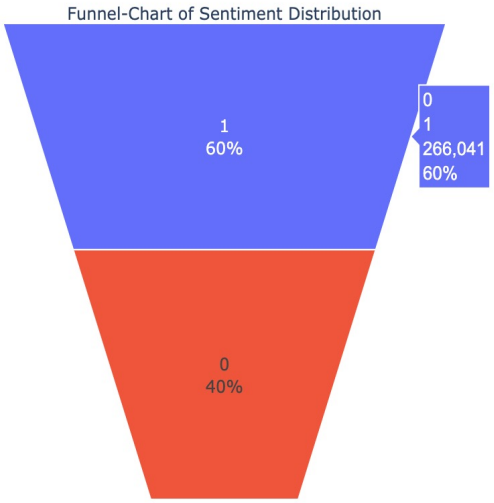
4
Project Result

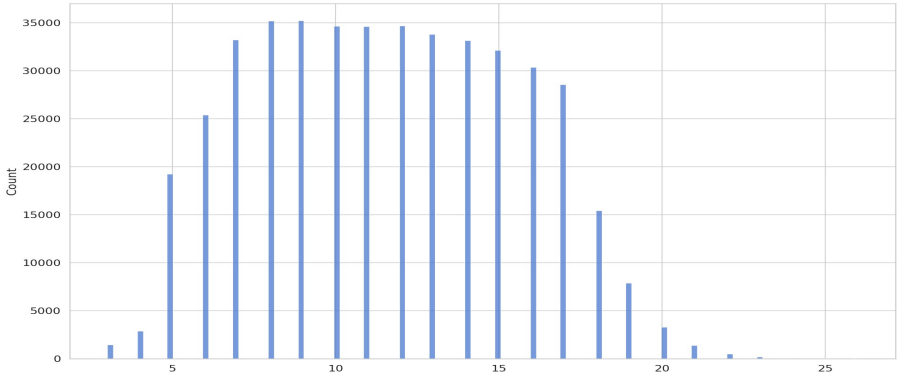
Project Result

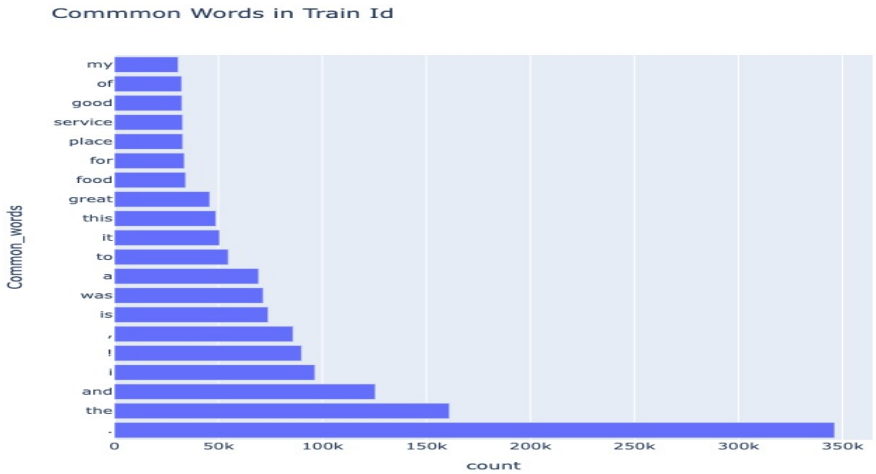
EDA

Train Data(Id, Category)

count	443259
Duplication	62689.



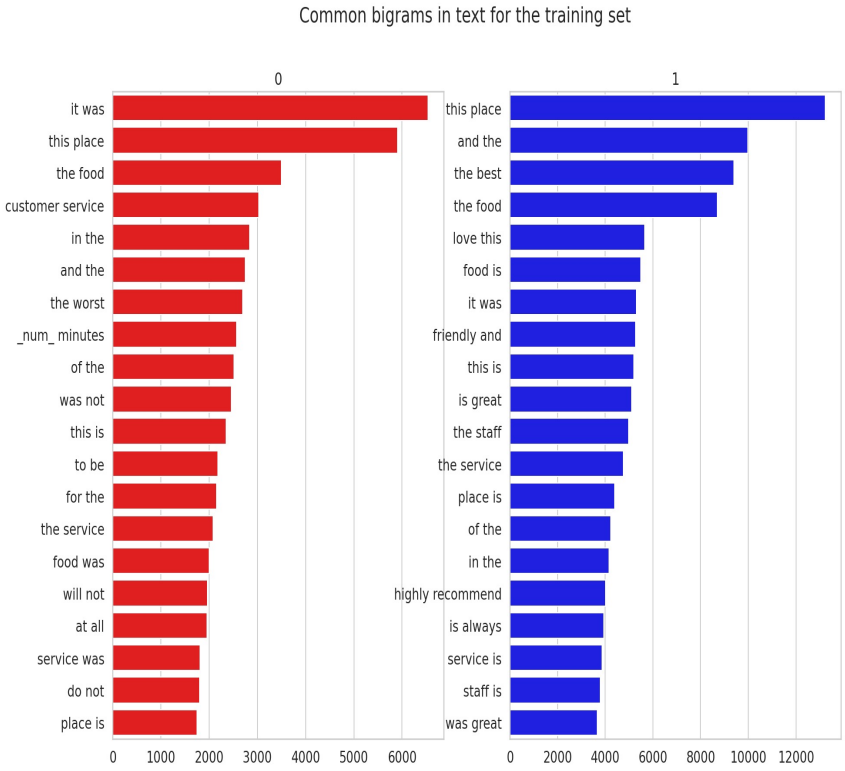


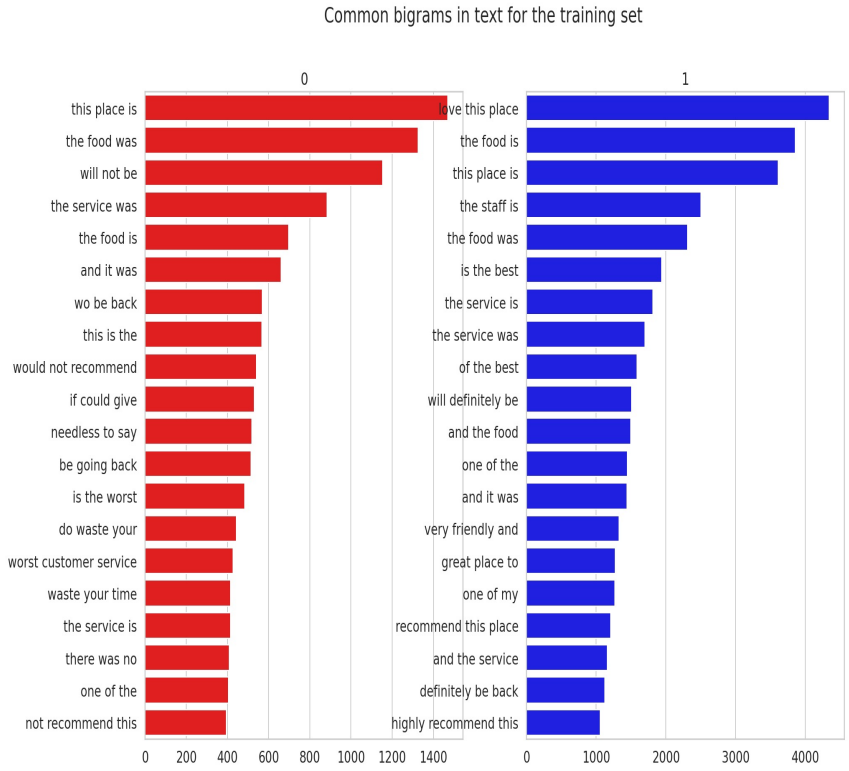


Project Result

EDA

Train Data(Id, Category)



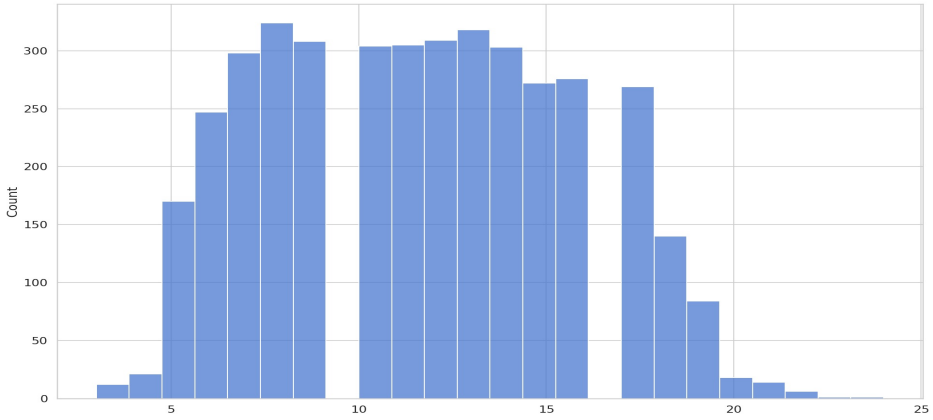


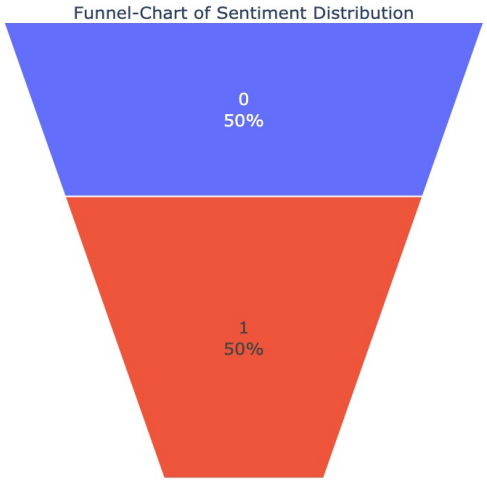
Project Result

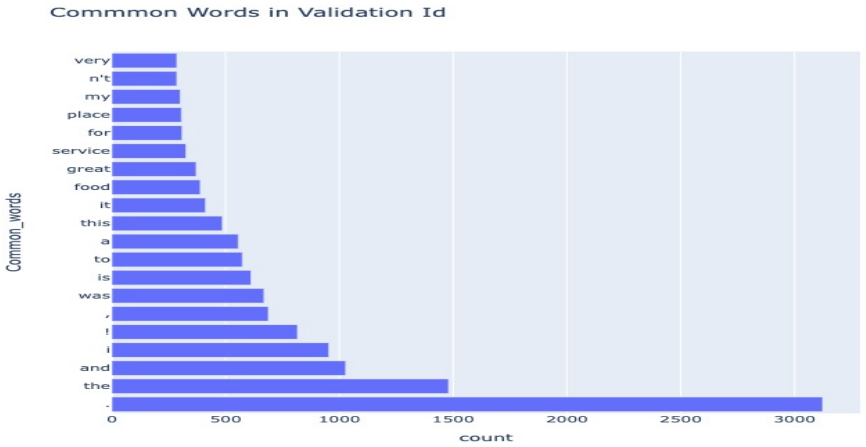
EDA

Validation Data(Id, Category)

count	4000
Duplication	145

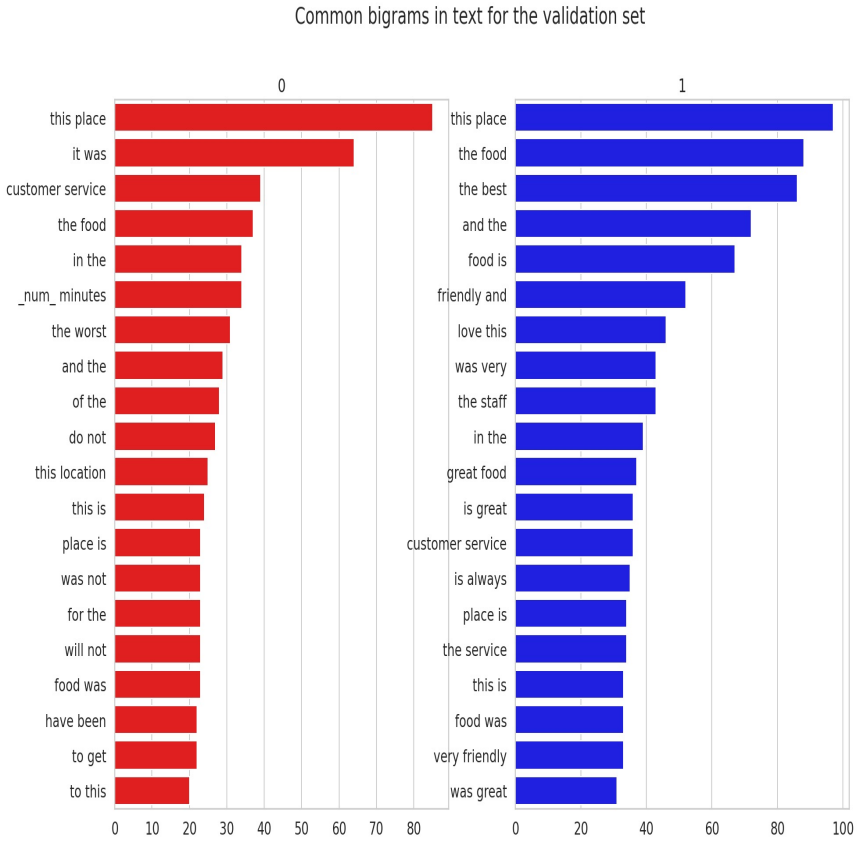


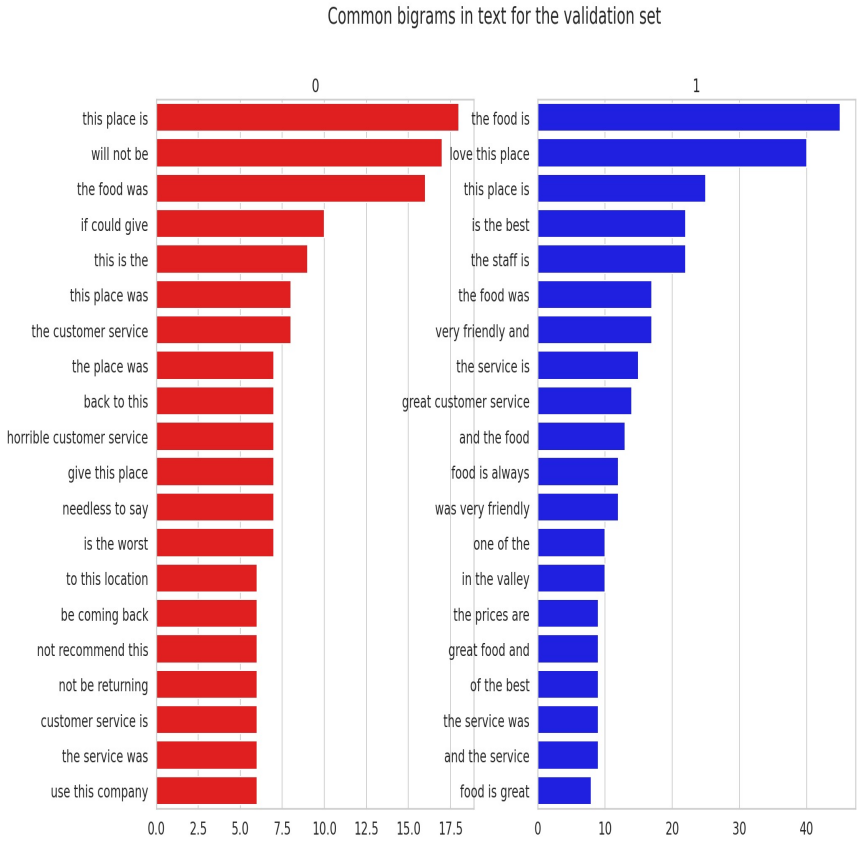




Project Result

EDA



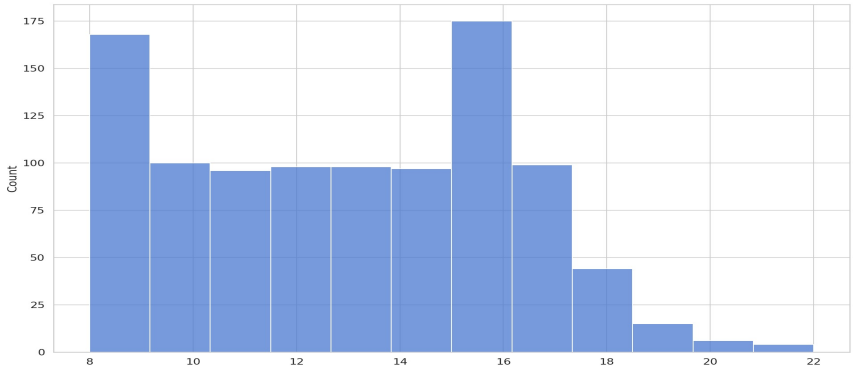


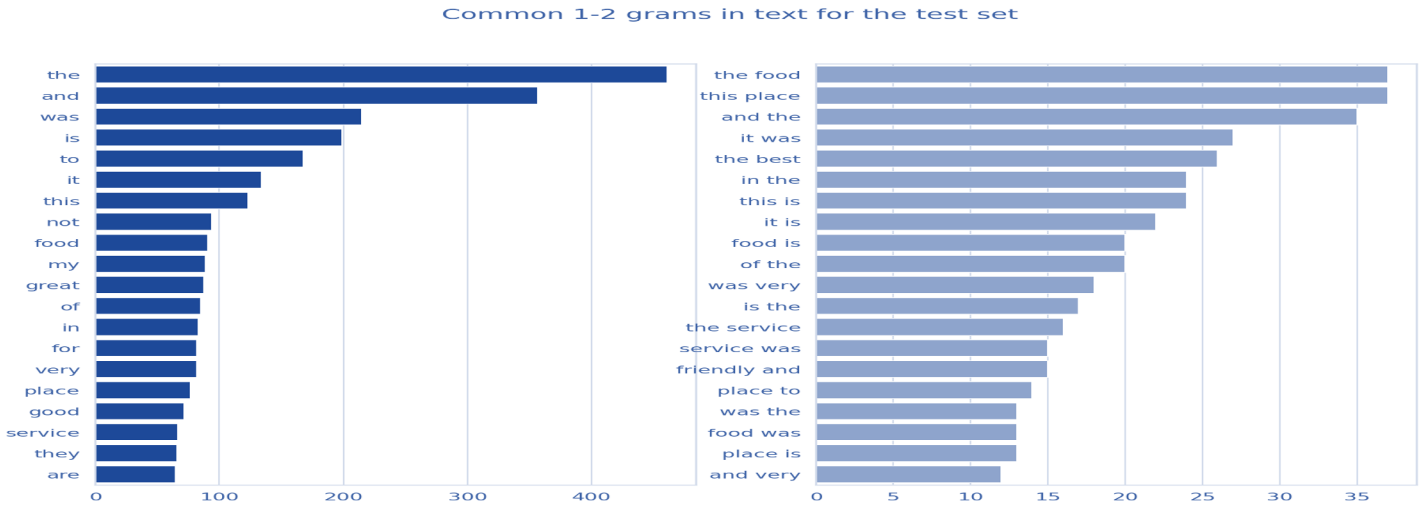
Project Result

EDA

Test Data(Id)

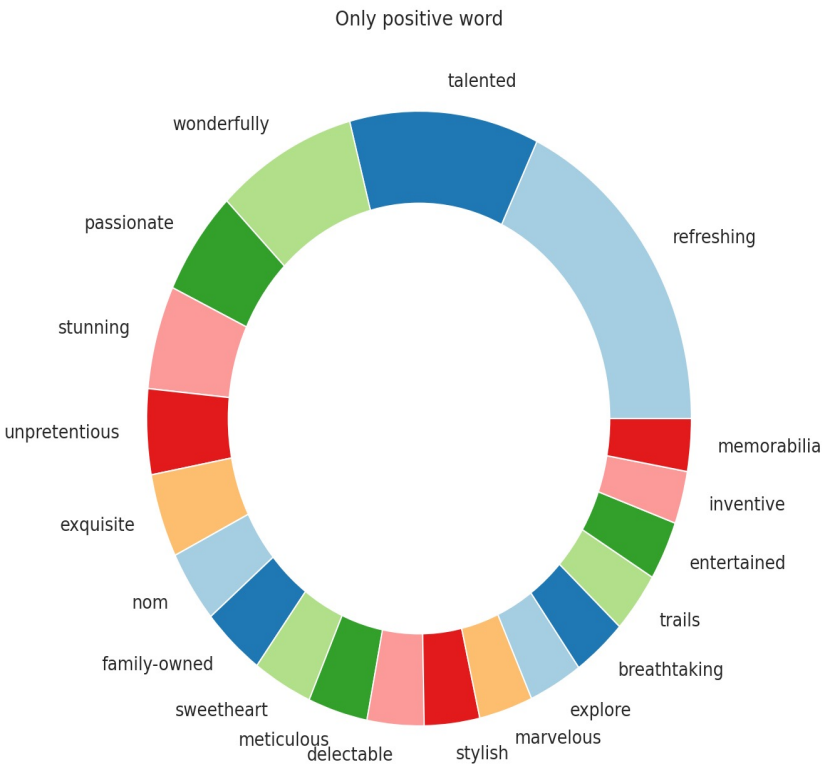
count	1000
Duplication	0

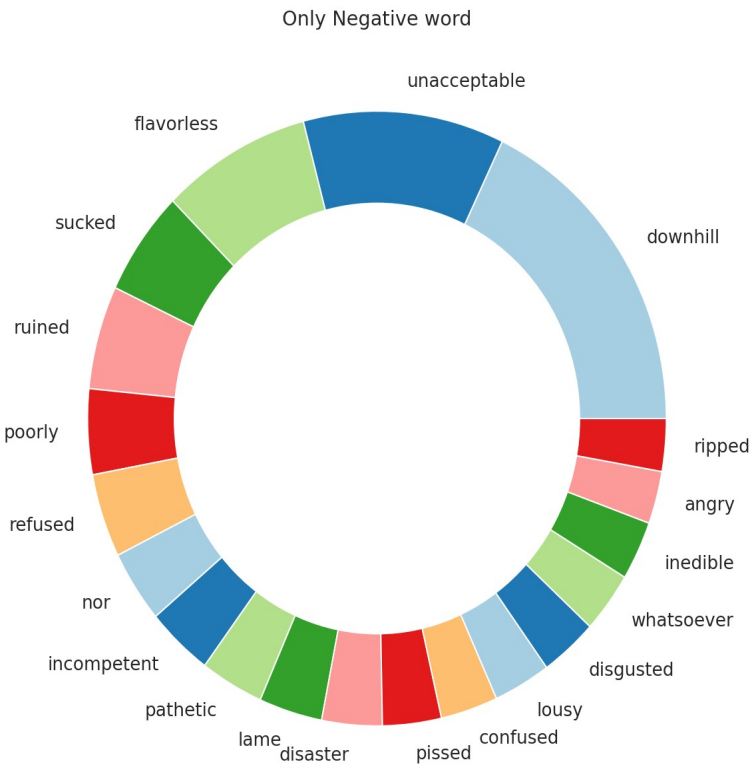




Project Result

EDA





Project Result

Preprocessing

1. 축약어 n’t -> not , ’s -> is, ‘m -> am, ‘ll -> will

축약어를 바꿔준 결과 정확도가 좋아짐.

2. 중복 제거

중복 제거를 통해서 6만개 정도의 데이터를 전처리한 결과 정확도가 낮아짐

3. 불용어 제거

불용어 제거 결과 정확도 낮아짐

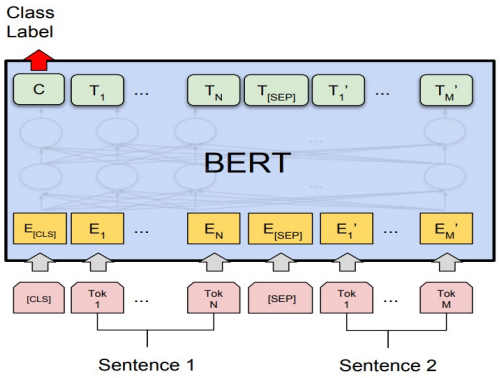
4. 특수문자 제거

!, $ 등의 특수문자 제거 결과 정확도가 좋아짐

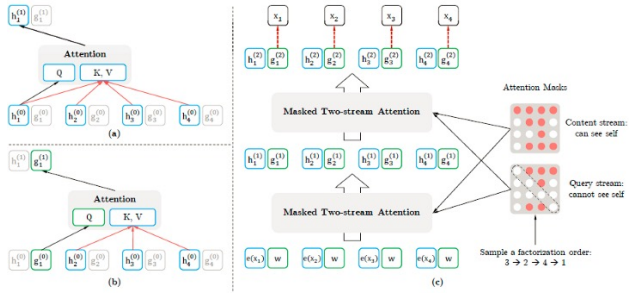
데이터 전처리는 **축약어**와 **특수문자 제거**만 진행한다.

Project Result

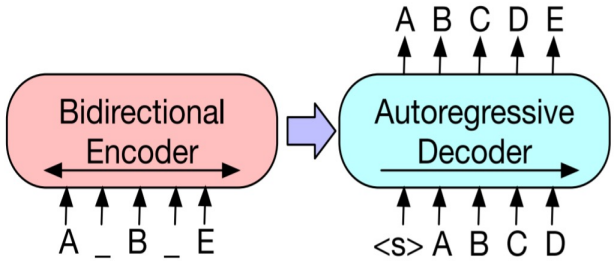
Model selection and analysis



BERT



XlNet



BART

Project Result

Model selection(BERT) and analysis

	Batch size	Weigth decay	Accuracy
Bert-base	64(tr)/64(eval)	-	0.979
Bert-base	64(tr)/64(eval)	(0.1)	0.984
Bert-base	256(tr)/256(eval)	(0.1)	0.983
Bert-base	256(tr)/256(eval)	(0.1)	0.980

Test Review

- 모델의 Batch size가 증가할 때 전체 accurac가 증가하는 경향을 보임
- 모델에게 weight decay를 적용할 시 accuracy가 증가하는 경향을 보임
- 둘 다 적용할 시 증가하는 추세를 보일 줄 알았으나 그렇지 않았음

Project Result

Model selection(xLnet) and analysis

Model	Batch size	Weigth decay	Epoch	Accuracy
xlnet-base	64	-	3	0.983
xlnet-base	64	-	5	0.983
xlnet-base	256	-	5	0.984
xlnet-base	256	0.9	5	0.986

Test Review

- 같은 batch size에서 epoch을 늘렸을 때, accuracy가 증가할 것이라고 예상하였으나 같은 accuracy를 보임
- batch size를 늘리자 accuracy가 증가함.
- Weight decay를 추가하였을 때, accuracy가 증가함.

Points of Improvement

- classification에 활용한 문장들의 경우 대부분이 길이가 짧아 XLNet이 가지는 긴 문장에 대한 이점을 활용하기가 어려웠다는 아쉬움이 있음.
- XLNet-large 모델을 사용하였을 경우, 성능 향상이 이루어질 수도 있다는 가능성을 고려할 수 있음.
- 다른 모델과의 앙상블 등을 통한 성능 향상 기대가 가능할 것으로 고려된다.

Project Result

Model selection(BART) and analysis

Model	Batch size	Weigth decay	Epoch	lr scheduler	Accuracy
Bart-base	64	0.1	2	liner	0.988
Bart-base	64	0.1	2	cosine	0.986
Bart-base	64	0.1	2	hard_restart	0.985
Bart-Large	64	0.1	1	linear	0.982

Test Review

- Bart-base 모델이 bart-Large 모델보다 성능이 좋음
- xLnet, bert 모델보다 성능이 좋다.
- Learning rate Scheduler 중 linear가 성능이 가장 높았다.

.

Project Result

Model selection(RoBERTa) and analysis

Model	Batch size	Weigth decay	Leaning rate	Epoch	Accuracy
roberta-base	64	0.1	0.00005	5	0.987
roberta-base	128	0.1	0.00005	5	0.988
roberta-Large	64	0.1	0.00005	5	0.989
roberta-base	256	0.1	0.00005	5	0.991

Test Review

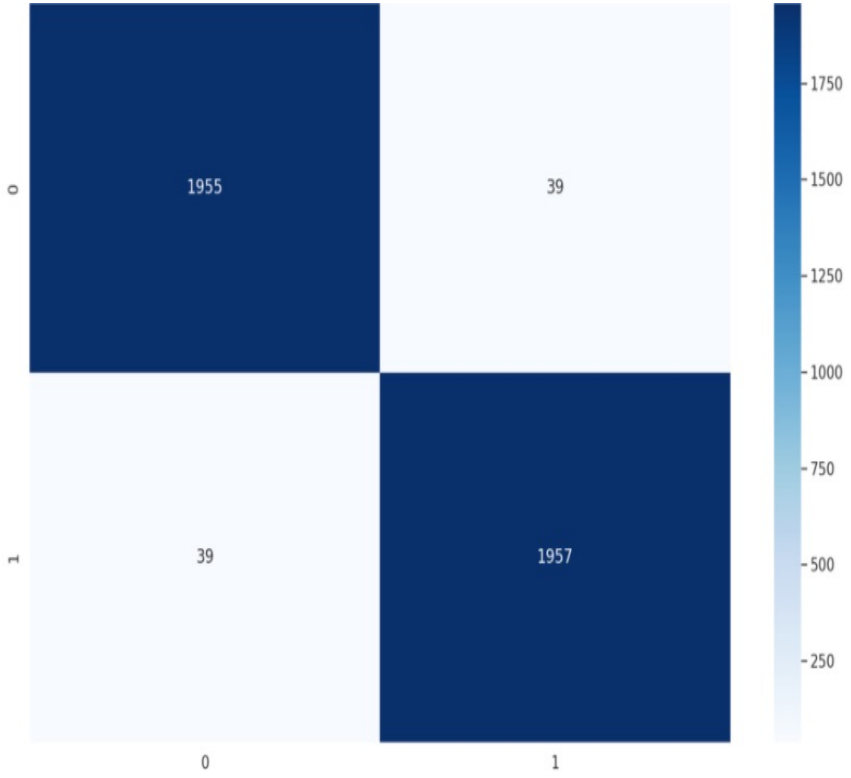
- Roberta-base model의 경우 batch size 를 증가했을 때 정확도가 증가함
- Roberta-large model이 Roberta-base model 보다 정확도가 높다..
- roberta-base의 batch size 를 256 으로 했을 때 가장 높은 정확도를 보임.

Points of Improvement

- GPU 메모리 부족으로 Roberta-large model에서 추가적인 batch size 조정을 하지못했다. Batch size의 크기를 늘리고, learnig rate 를 다양한 값으로 시도해 볼 수 있다.

Project Result

Model selection(RoBERTa) and analysis



Negative -> Positive

she rings me up and hands me one cup

i keep trying to give this place a chance

Positive -> Negative

for a moment i thought i m eating mom food

num bucks

5
자체 평가 및 보완

자체 평가 및 보완

다양한 모델 활용

다양한 모델을 문제에 적용해 보면서 각 모델의 특성에 대해서 파악하고 문제에 적합한 최적의 모델을 찾는 경험을 했다.

Gpu의 메모리 용량 제한

Gpu의 메모리 용량의 한계로 input data의 max_length의 길이를 늘리는데 한계가 있고, Batch_size의 크기에도 제한 사항이 있다.

HyperParameter Tuning

Learning_rate, epoch, Regularizaztion 등의 파라미터 튜닝을 다양하게 진행하지 못했다.

Esemble 적용

Esemble 을 적용하여 모델의 정확도를 높이려는 시도를 해볼 수 있다.

THANK YOU!