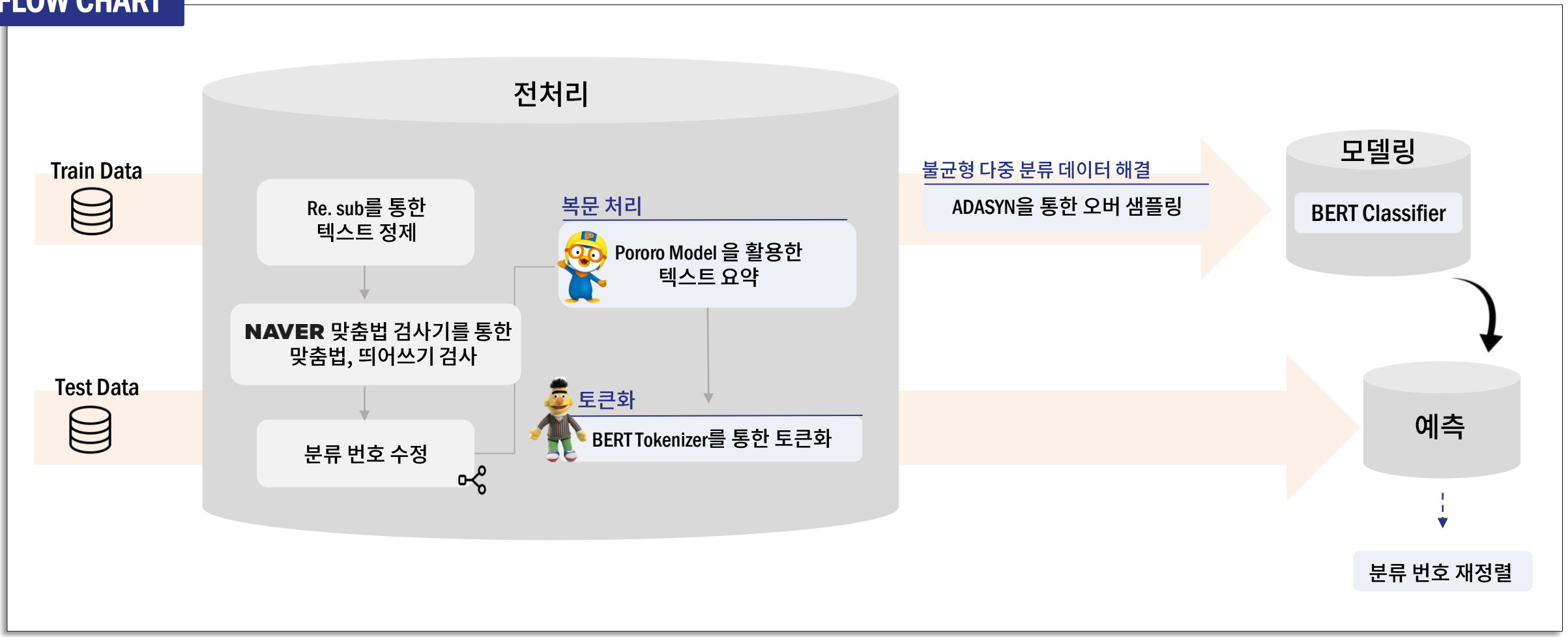
2021 SCIC 삼성카드 데이터 분석 & 아이디어 공모전

[Track1] 고객 피드백 분류 모델 기획서

팀명 E - poch 팀원 서혜정, 예지혜, 최보금

FLOW CHART



데이터 전처리

TEXT CLEANING 텍스트 정제

Package: Re

파이썬은 정규 표현식을 지원하기 위해 re(regular expression) 모듈을 제공한다.

re.sub은 패턴에 일치되는 문자열을 다른 문자열로 바꿔주는 함수이다.

❖ 사용 이유

갖고 있는 말뭉치(corpus)로부터 불필요한 문자(예: 특수문자 등)를 제거한다.

<코드 예시>

```
sent = df['발화'][679]

def clean_text(sent):
    sent_clean=re.sub("[^가-힣ㄱ-ㅎㅏ-ㅣA-Za-z1-9]", " ", sent)
    return sent_clean

print(sent); print(clean_text(sent))
```

```
>>> 친절한 상담 감사합니다~~^^
>>> 친절한 상담 감사합니다
```

NAVER SPELL CHECKER 네이버 맞춤법 검사기

Package: Hanspell

네이버 맞춤법 검사기를 이용한 파이썬용 한글 맞춤법 검사 라이브러리이다.

검사기를 통해 확인된 맞춤법 오류, 띄어쓰기 오류, 통계적 오류, 표준어 의심 텍스트를 교정 결과로 대치해준다.

❖ 사용 이유

사용자 리뷰 특성상 맞춤법 오류로 인해, 동일한 내용을 다르게 인식하는 문제가 발생할 수 있다. 맞춤법 교정은 이러한 문제를 해결하여 모델의 학습 성능을 높여준다.

<코드 예시>

```
sent = df['발화'][201]
print(sent)

spelled_sent = spell_checker.check(sent)
hanspell_sent = spelled_sent.checked
print(hanspell_sent)
```

```
>>> 카드만들기가힘들엇다  친절이조금고푸다
>>> 카드 만들기가 힘들었다  친절이 조금 고프다
```

CLASS NUMBER EDIT 분류 번호 수정

❖ 사용 이유

발화가 분류되어 있지 않은 클래스를 지우고 버트를 실행하기 위함이다.

TEXT SUMMARIZATION 텍스트 요약

Package: Pororo (Bullet point summarization)

Pororo는 카카오브레인(Kakao brain)에서 공개한 오픈소스로서 다양한 자연어 태스크에 대응 가능한 통합된 형태의 자연어 프레임워크이다. KoBart 모델을 베이스로 하며, 한국어 태스크들에 대해 좀 더 최적화 되어 있다는 장점을 지닌다.

Pororo는 3가지의 서로 다른 summarization 기능(Abstractive, Extractive, Bullet) 을 제공한다. 세 방법 중 가장 사용자 리뷰를 잘 요약해준 Bullet을 요약 모델로 선택하였다.

❖ 사용 이유

갖고 있는 말뭉치(corpus)로부터 불필요한 문자(예: 특수문자 등)를 제거한다.

- 복문 가능성을 판단하는 기준: 문장 길이가 85자 이상인 리뷰

- 해당 기준을 정한 이유

1000개의 샘플 데이터에서 제공된 복문의 갯수는 총 32개였다. 상위 3.2%에 해당하는 발화 길이는 85자이다. 85자를 기준으로 그 이상부터 복문일 가능성이 높다고 판단했다.

<코드 예시>

```
summ = Pororo(task="summarization", model="bullet", lang="ko")
print(df['발화'][311])
result = summ(df['발화'][311])
print(''.join(result))

>>> 좋았습니다. 다만 보험 전화 좀 그만하셨으면 좋겠습니다. 이미 가입한 보험이 있는데도 삼성카드로 연락 자주 오고 있습니다.
있다고 받았는데도 집요하게 연락 와서 힘들었습니다.
>>> 보험 전화 그만하셨으면 좋겠다 이미 가입한 보험인데 삼성카드로 연락 자주 와
```

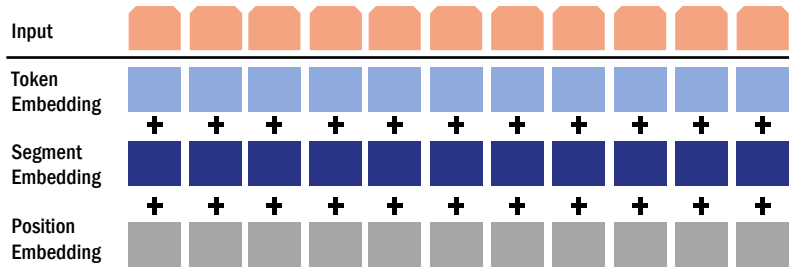
토큰화

BERT TOKENIZER 버트 토크나이저

Method: BERT tokenizer

3가지의 입력 임베딩(Token, Segment, Position 임베딩)의 합으로 구성된다.

Token Embedding(Word Piece 임베딩)은 모든 언어에 적용 가능하며 sub-word 단위로 단어를 분절하므로 Out Of Vocabulary 처리에 효과적이고 정확도 상승효과도 있다. 즉, 희귀 단어 · 이름 · 숫자나 단어장에 없는 단어에 대한 학습과 번역에 수월하다.



예시

Input	“안내원 연결까지 대기시간이 너무 길어요” Input에 대한 Embedding 결과
Token Embedding	[101 9521 31605 14279 9568 74322 18382 9069 12310 100699 10739 9004 32537 8934 12965 48549 102 0 0 0 • • • 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
Segment Embedding	[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 • • • 0 0 0 0]
Position Embedding	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 • • • 0 0 0 0]

불균형 데이터 처리

OVERSAMPLING 오버 샘플링

Method: ADASYN(adpative synthetic sampling approach)

위치에 따라 다르게 SMOTE 적용하는 방식으로 각 소수 클래스 주변에 얼만큼 많은 다수 클래스 관측치가 있는지 정량화 한다.

1) 모든 소수 클래스에 대해 주변의 k개 만큼 탐색하고 그 중 다수 클래스 관측치의 비율 r_i을 계산한 후 표준화 한다.

2) 표준화된 r_i 를 생성하고자 하는 개수 (G = 다수클래스 개수 - 소수 클래스 개수) 만큼 곱한다.

3) r_i 의 정도에 따라 생성되는 샘플의 수가 다르다.

Δ_i : 소수 클래스 xi의 주변 k 개중 다수 클래스의 관측치 개수

m : 소수 클래스 내 관측치 총 개수

$$r_i = \frac{\Delta_i}{K} \quad i = 1,2,3,\ldots,m$$

범주	데이터 수
0	207
1	27
2	128
3	97
4	3
5	5
6	10
7	97
8	15
9	24
10	5
11	100
12	8
13	2
14	4
15	7

ADASYN 적용

데이터 수
207
202
232
219
207
205
210
173
211
211
205
182
208
206
208
210

모델링

BERT MODELING 버트 모델링

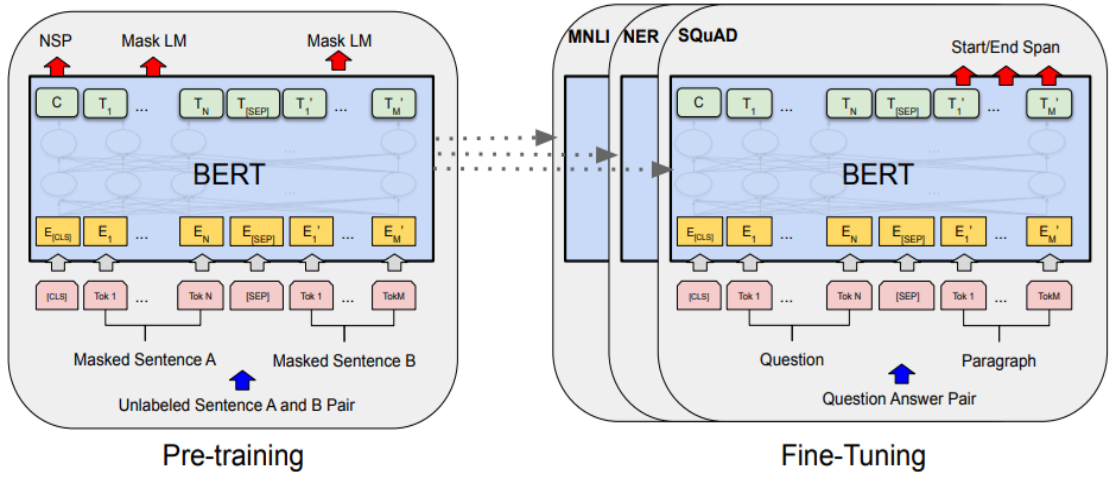
❖ 사용 이유

리뷰 텍스트 데이터 셋의 특성상 사용자에 따라 다양한 표현방식이 존재한다. 학습의 정확도를 높이기 위해선 발화의 문맥을 고려하는 것이 필수적이다. Bert 모델은 AI 최첨단 딥러닝 모델로서 각 단어의 양방향에 위치한 단어들을 통해 문맥이 고려되기 때문에 해당 모델을 선택하였다.

❖ 알고리즘 설명

Bert 모델은 크게 사전 학습(Pre-training) 과정과 미세 조정(Fine-Tuning) 과정으로 나뉜다. 마스킹된 토큰을 예측하는 MLM 방식과 문장의 연관성을 예측하는 NSP 방식으로 사전 학습된 모델을 로드하여, 리뷰 데이터로 미세 조정 학습을 시켰다. 학습시킨 모델에 분류하고자 하는 데이터를 입력하여 결과값을 얻고, 완전 연결층을 활용해 예측 클래스를 출력하는 구조이다.

학습 인풋	학습 데이터의 토큰 (버트 토크나이저 결과) 학습 데이터 라벨
파라미터	max length of sentence batch size epoch validation split
예측 인풋	예측 데이터의 토큰 (버트 토크나이저 결과)



[출처] 버트를 파헤쳐 보자!! https://keep-steady.tistory.com/19

후처리

CLASS NUMBER REMATCH 분류 번호 재정렬

❖ 사용 이유

버트를 실행하기 위해 수정했던 분류 번호를 재매칭하여 예측 결과의 실제 클래스를 찾는다.

	발화	의도명(유형)
0	고객 입장에서 잘 듣고 요점만 정확하게 답변했으면 좋겠습니다	불만>고객서비스>상담원
1	친절하게 상담해주어서 감사합니다	칭찬>고객서비스>상담원
...	...	...
245	친절하게 해주셔서 감사합니다	칭찬>고객서비스>상담원
246	4번 5번 문항 답변 불편한 게 없었는데 체크하라고 강요합니다 상담원분들 더우신데 ...	불만>고객서비스>상담원