삼성카드

목차

1. 선정 알고리즘
2. 데이터 전처리
3. 후처리 및 모의 분석

"고객 피드백 분류"를 위한 알고리즘 모델 기획서

Team : 머홍

1 선정 알고리즘

대분류	중분류 및 소분류
LSTM or GRU	KoBERT

1. LSTM, GRU

*비교적 가벼운 언어모델

➢ **무거운 Kobert의 단점을 상쇄시키기 위해 사용**

2. Kobert[1]

*구글에서 개발한 언어모델인 BERT의 한국어 버전

* 현재 가장 보편적이며 성능이 좋은 알고리즘

* Inference가 느리다는 단점

➢ **높은 분류 정확도를 위해 사용**

Idea Challenge의 objective

1. 다중분류의 정확도
2. 가볍고 성능이 좋은 모델

문제 해결 Process

1. LSTM or GRU를 통해 '칭찬'과 '불만' 분류

*** '중립' 과 '폐기'는 비교적 데이터가 적을거라 예상**

(1)중립 해결방안

➢ 동사가 없거나 명사구로 되어있기 때문에 간단한 조건문으로 해결

(2)폐기 해결방안

➢ anomaly detection 또는 Rejection Option[2]을 통해 모델의 output score가 threshold내이면 폐기로 분류

2.Kobert로 중분류와 소분류를 합쳐서 진행

*** 중분류와 소분류를 합쳐 총 21개의 카테고리**

ex) '고객서비스-상담원', '삼성카드-커뮤니케이션'

[1] https://github.com/SKTBrain/KoBERT

[2] Recent Advances in Open Set Recognition: A Survey [Chuanxing Geng *et al., IEEE 2020]*

2 데이터 전처리 - 1

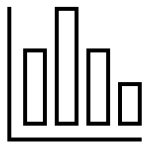
1. 맞춤법 검사

- Py-Hanspell

✓ 네이버 한글 맞춤법 검사기를 바탕으로 만들어진 패키지

✓ Training시 적용하면 모델이 맞춤법이 맞지 않는 발화에 robust하지 못함

➢ **Inference시에만 적용**



- 제공 데이터 실습 결과

1) Train 데이터에 맞춤법 검사를 하고 학습시킨 모델

2) Train 데이터에 맞춤법 검사를 하지 않고 학습시킨 모델

* Test data accuracy 결과

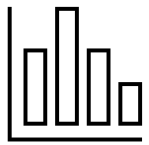
2 > 1

➢ **Training시 Py-Hanspell 미적용**

2. 토큰화

- 형태소 분석기 종류

1. Mecab
2. KoNLPY의 Okt
3. Pretrained된 문장에 KoBERT의 tokenizer 적용



- 제공 데이터 실습 결과

1) Mecab

2) Okt

3) KoBERT의 tokenizer

* Test data accuracy 결과(같은 조건)

1 > 2 = 3

➢ **Mecab을 사용하여 토큰화**

3. 복문 처리

- 복문 처리 방법

✓ 복문 또는 구형태의 문장은 주로 접속사 이후에 발화의 목적이 담겨있음

➢ **토큰화시 접속사등을 기준으로 뒷문장에 집중**

2 데이터 전처리 - 2

4. 대분류

- Labeling

✓ '중립' 또는 '폐기' 분류는 조건문과 rejection option 또는 anomaly detection으로 해결

➢ **'대분류'는 '칭찬'과 '불만'의 binary classification**

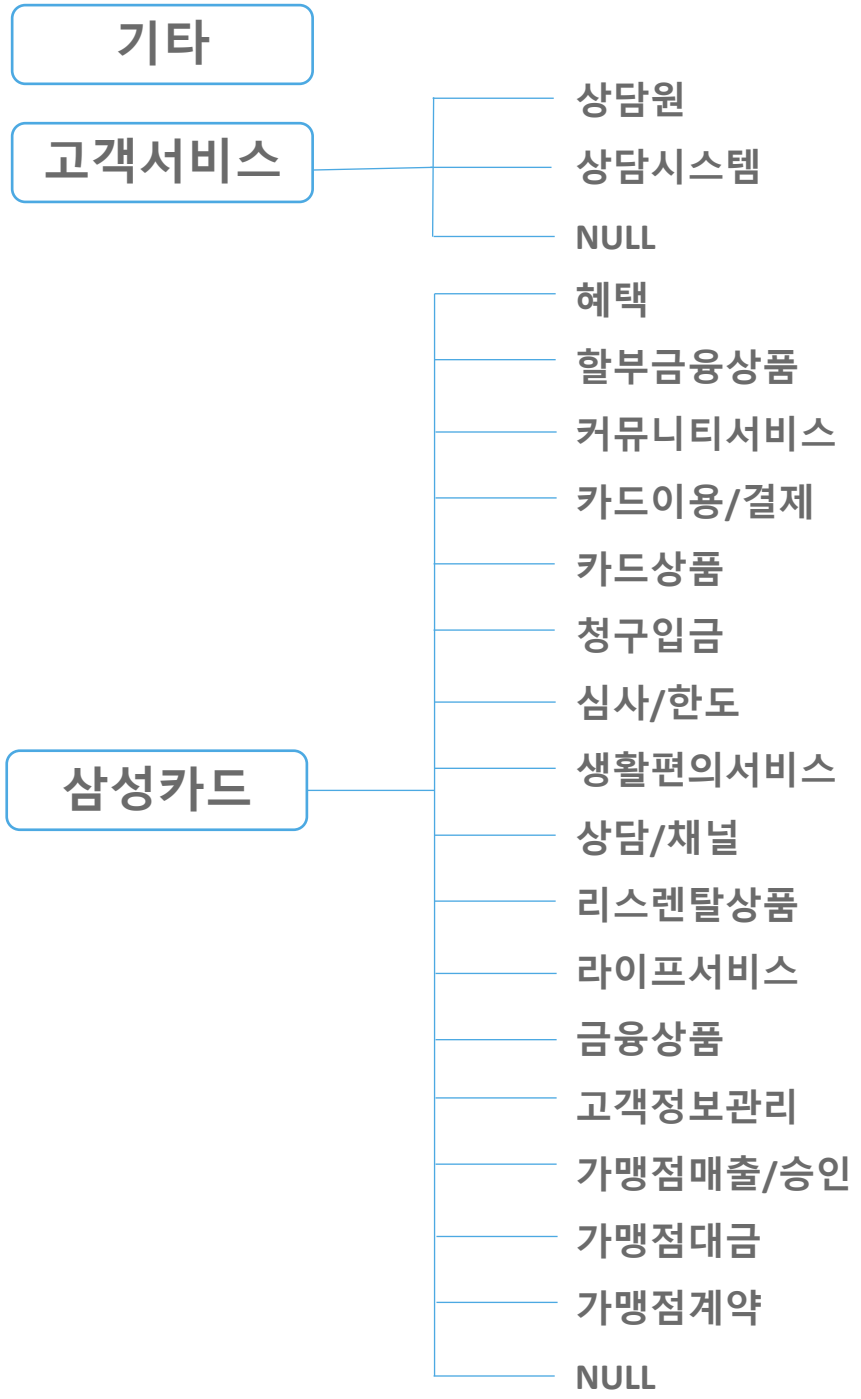
5. 중,소 분류

- Labeling

✓ 중분류는 '삼성카드', '고객서비스', 기타' 3가지

✓ 삼성카드의 소분류 17가지, 고객서비스의 소분류 3가지, 기타를 합쳐 총 21가지로 분류

➢ **총 21가지의 카테고리로 나눈 multi classification**



3 후처리 및 모의 분석

후처리 방법

- 결과값 처리

✓ 제공 데이터에서 복수개의 분류가 채택된 발화가 존재함을 확인

✓ 복수 분류의 경우 우선순위를 통해 최종 분류가 되는 조건이 있음을 확인

➢ **Category별 %와 우선순위 score를 지정하여 최종분류의 근거를 보여줄 계획**

모의 분석

- LSTM 모델을 이용한 대분류

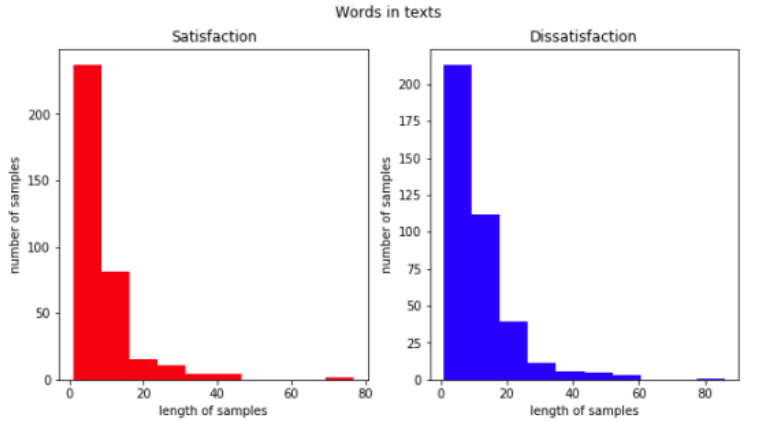
```
전체 데이터 수: 1000
모델이 맞춘 데이터 수: 956
모델이 틀린 데이터 수: 37
정답률: 0.956
1000개의 데이터에 대한 수행시간: 29.017874717712402
```

- KoBERT 모델을 이용한 대분류

```
전체 데이터 수: 1000
모델이 맞춘 데이터 수: 987
모델이 틀린 데이터 수: 13
정답률: 0.987
1000개의 데이터에 대한 수행시간: 180.29840993881226
```

✓ LSTM과 KoBERT를 통해 Inference 시간과 정확도가 trade-off 관계인것을 알수 있었으며 실험을 통해 본 challange의 GOAL에 부합한 알고리즘을 설계

```
칭찬 발화의 평균 길이 : 8.694915254237289
불만 발화의 평균 길이 : 11.923076923076923
```




```
발화의 최대 길이 : 61
발화의 평균 길이 : 9.065860215053764
```