



ChatGPT 생성 리뷰를 활용한 콘텐츠 평가

기계학습특론 3조

김현준 이윤지

목 차

1. 연구 개요
2. 방법론
 - 1) 데이터 수집
 - 2) 데이터 전처리
 - 3) 모델 학습
3. 실험
 - 1) 모델 성능
 - 2) App 별점 예측
 - 3) Youtube 별점 예측
4. 결론



1. 연구 개요

연구 배경

- 온라인 상의 다양한 콘텐츠에 대한 사용자의 반응 파악
- 점수화되지 않은 콘텐츠의 경우, 사용자의 반응을 빠르게 파악하기 어려움



연구 주제

ChatGPT로 생성한 리뷰 데이터 기반 콘텐츠의 사용자 반응 파악



연구 목표

- ChatGPT를 활용하여 데이터 생성 및 모델 학습
- 실제 환경의 콘텐츠에 대한 사용자 반응 계량화 및 예측

1. 연구 개요

주요 기여

- ChatGPT를 활용하여 독자적인 데이터셋 구축
- Naive Bayes, Random Forest, XGBoost 그리고 Ensemble 모델에서 데이터셋 검증
- 학습한 모델들을 활용하여 실제 환경에서 사용자 반응 파악 가능 여부 확인
 - (1) 4개의 Google Play store App 콘텐츠 적용
 - (2) 4개의 Youtube 동영상 콘텐츠 적용

2. 방법론

데이터 수집

ChatGPT를 활용한 데이터 생성 및 수집

데이터 전처리

독립 변수: 수집한 데이터의 텍스트에서 feature를 추출

종속 변수: 수집한 데이터의 라벨(별점)을 이진분류 형태로 변환

모델 학습

전처리 단계에서 추출한 feature들을 이용한 모델 학습

2. 방법론


1) 데이터 수집

- ChatGPT에 쿼리(query)를 요청하여 csv파일 생성
- 데이터 형태는 'contents', 'star' 컬럼으로 구성된 24,599개의 데이터

GPT-4

h

서비스 평가 댓글을 생성해 줘. 그리고 별점은 1~5점 사이로 평가해주고 부정적에 가까우면 1, 긍정에 가까우면 5로 해 줘.(균일한 비율로 생성) 그리고 컬럼은 **contents**, **star**로 해 줘. **contents**에 댓글 내용, **star**에는 별점을 넣어줘. 댓글 길이는 최대 100단어로 구성해서 랜덤하게 생성해 줘. 이런 조건으로 "csv 포맷", 50개 데이터 만들어 생성해 줘. 언어는 영어로 만들어 줘



물론입니다. 아래는 요청하신 대로 생성한 서비스 평가 댓글 50개의 예시 목록입니다. 댓글은 영어로 되어 있으며, 별점은 1에서 5 사이이고, 각 별점에 대해 균일한 분포를 가지고 있습니다.

CSV

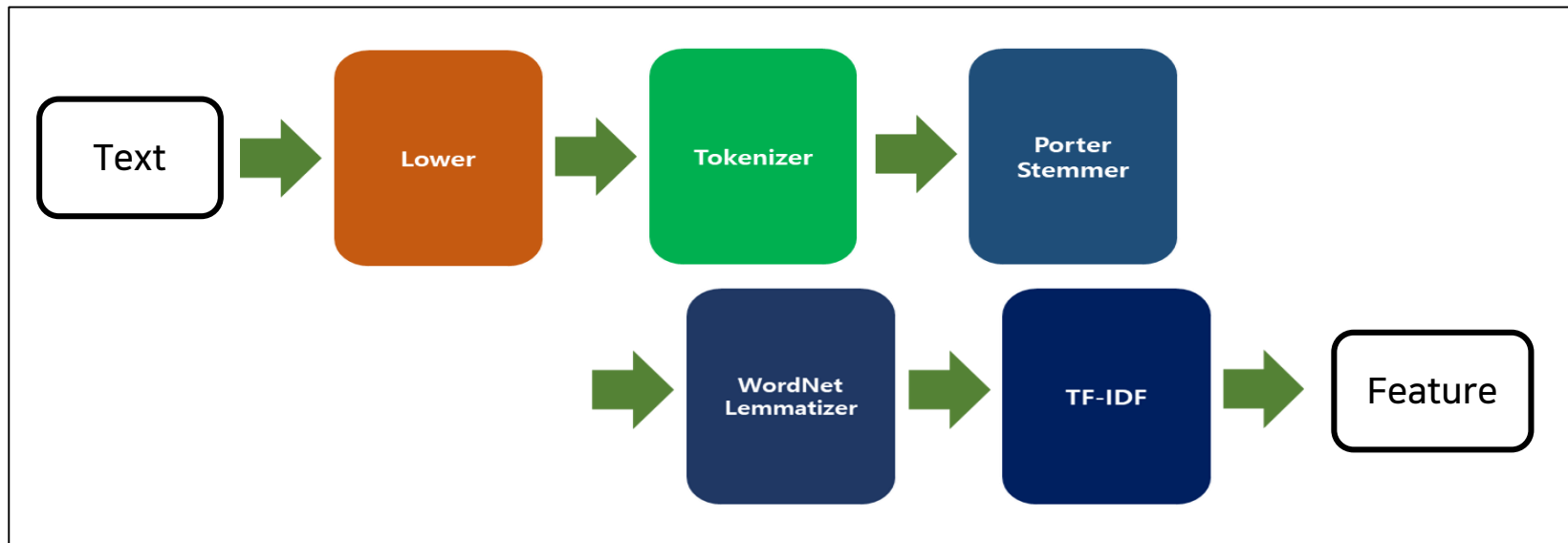
Copy code

```
contents,star
"Terrible service, waited too long and staff was rude.",1
"Quite disappointed. The product broke within a week.",1
"Service was mediocre. Not what I expected for the price.",2
"Food was okay, but the ambiance lacked warmth.",2
"Average experience. Nothing exceptional, but not too bad.",3
```

2. 방법론

2) 데이터 전처리 - 독립 변수

- 수집한 데이터를 아래와 같은 파이프라인(pipeline)을 거쳐 feature 추출
- 본 연구에서는 TF-IDF를 활용하여 텍스트의 단어 중요도를 계산하여 feature 추출



2. 방법론

2) 데이터 전처리 - 종속 변수

- ChatGPT에서 수집한 리뷰 데이터는 5개의 레벨(1~5)로 구성
- 레벨이 낮을수록 부정적인 내용, 높을수록 긍정적인 내용 의미
- 본 연구는 이진 분류를 목표로 하기 때문에 아래와 같이 라벨 조정 진행



* 레벨 3의 경우, 중립적인 표현을 포함한 데이터가 섞여 있어
본 연구에서는 명확한 분류를 위해 삭제함.

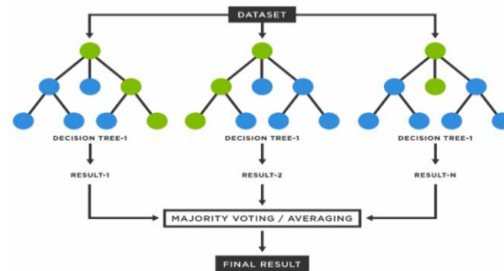
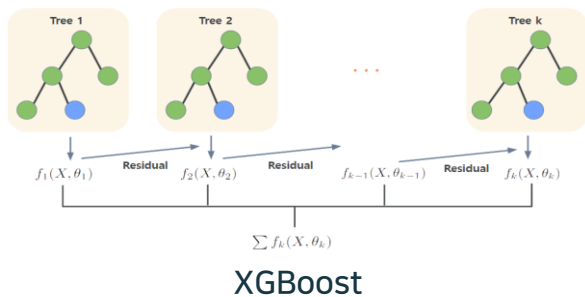
2. 방법론

3) 모델 학습

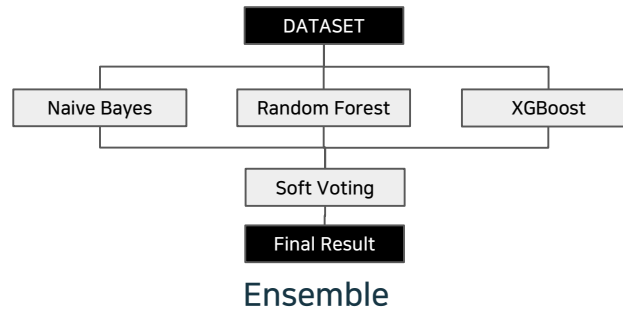
- 전처리를 거친 데이터를 8:2의 비율로 Train Set과 Test Set을 나누어 모델 학습
- 선정 모델: Naive Bayes, Random Forest, XGBoost, Ensemble

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Naïve Bayes



Random Forest



3. 실험

1) 모델 성능: ChatGPT 리뷰 데이터셋

	Train Accuracy	Test Accuracy
Naïve Bayes	93.17%	92.05%
Random Forest	99.77%	92.75%
XGBoost	92.83%	90.96%
Ensemble	97.83%	93.91%

3. 실험

2) App 별점 예측

	ChatGPT	Viking Rise	One Star App	Bin Fine Opener
Actual	4.4	4.4	1.5	1.7
NB	3.84	3.97	2.02	1.76
RF	3.47	3.53	2.2	1.46
XGB	3.55	3.73	2.2	1.71
Ensemble	3.72	3.86	1.99	1.58

3. 실험

3) Youtube 별점 예측

- Video1과 Video2는 교육용 강의 콘텐츠 → 긍정적 결과 예상
- Video3과 Video4는 전쟁 뉴스 콘텐츠 → 부정적 결과 예상

	Video1	Video2	Video3	Video4
NB	3.74	4.42	2.49	2.33
RF	3.14	4.42	1.82	1.79
XGB	2.98	4.42	1.77	1.86
Ensemble	3.32	4.71	1.98	1.92

4. 결론

의의

- 새로운 가상 데이터를 생성함으로써 연구 목적에 맞는 데이터 확보
- 학습한 모델을 실재하는 외부 콘텐츠에 적용하여 평가함으로써 모델의 실용성 및 활용성 확보
- Train set과 Test set을 분리하여 학습을 검증함으로써 데이터의 일관성과 무결성 확보
- Ensemble 모델의 경우, 모델들 간의 장·단점 보완

한계

- 텍스트 분석은 종립적 표현과 반어법 사용 등의 이유로 해석이 모호한 경우가 빈번하게 발생
- 1차적으로 해석이 모호한 데이터를 필터링한 후, 본 연구 내용을 적용한다면 더 나은 결과를 도출할 것으로 예상