**1. 프로젝트 개요**

최근 eCommerce 시장이 발달하면서 리뷰 데이터가 점점 더 중요하게 여겨지고 있다. 직접 보고 만져볼 수 없는 온라인 쇼핑의 특성상 미리 제품을 사용해 본 소비자의 의견은 타 소비자들 입장에서는 제품을 구매하는 데에 중요한 지표가 되며 기업 입장에서는 제품의 브랜드를 Build-up 해나갈 수 있는 자산이 된다. 특히 리뷰의 사이즈와 긍부정 지수는 많은 연구를 통해 매출과 상관관계가 있다고 하며 오늘날의 많은 기업들은 소비자의 긍정적인 리뷰를 이끌어내기 위해 끊임없이 연구함과 동시에 부정적인 리뷰를 분석하여 문제가 있는 제품들을 Renewal 해나가는 형태로 변해가고 있다.

하지만 현재까지 한국어 특성상 문맥 분석이 어려워 낮은 긍부정 지수 정확도를 기록하고 있으며 지금까지도 보다 높은 수준의 한국어 자연어처리 모델 개발을 위해 많은 연구가 이뤄지고 있다. 이 프로젝트에서는 BERT라는 사전학습 방식의 딥러닝 기법을 통해 소비자의 리뷰 데이터를 학습하여 80% 이상의 긍부정 정확도를 보여주는 모델을 개발해보려고 한다.

**2. 프로젝트 목적**

해당 프로젝트는 가상 시나리오 기반으로 문제해결을 위해 거쳐 온 단계를 설명한다.

- 제조기업 ICT는 최근 제품의 Market Share가 떨어진다는 report를 받았다

- ICT 재무팀은 데이터팀에 의뢰하여 매출이 떨어지고 있는 이유를 알아봐달라고 했다.

- 데이터팀은 고객 리뷰데이터를 통해 현재 ICT 제품에 대해 보다 빠르게 고객들의 긍부정 지수를 추출해주는 모델을 개발하려고 한다.

**3. 프로젝트 내용**

해당 프로젝트는 BERT를 통해 새롭게 작성 된 리뷰 데이터의 긍부정 지수는 어느 정도인지 판단해주는 자연어처리 모델 개발 과정을 담고 있다.

**3-1. 사용 데이터셋**

Coupang review data (2019~2022)

쿠팡 내에 소비자가 작성한 리뷰 데이터를 크롤링하여 사용함.

**3-2. 데이터 전처리**

Data Description

date: 리뷰 작성 날짜

year: 리뷰 작성 연

month: 리뷰 작성 월

day: 리뷰 작성 일

brand: 제품 브랜드

vid: 제품 ID

subc: 제품 서브브랜드

type: 제품 타입

size: 제품 사이즈

iteam\_name: 제품명

review: 리뷰 원문

rating: 리뷰 평점

**Feature Engineering**

평점을 기준으로 1-3: 부정[0] / 4-5: 긍정[1] 인 target (label) 컬럼을 추가함.



A picture containing text, screenshot, monitor, black

Description automatically generated

그 후 분석에 필요한 컬럼들만 추출.



Graphical user interface, text, application

Description automatically generated

**3-3. EDA**

Label 분포를 확인해 본 결과 굉장히 불균등한 분포를 가지고 있었음.

\* 불균등할 경우 모델의 성능이 나빠질 수 있음

\*\* Startify를 통해 불균형을 보정할 수 있지만 데이터셋 수가 너무 적어지게 되어 일단은 그냥 진행하도록 하겠습니다.

Chart

Description automatically generated with medium confidence

**3-4. 데이터 분리**

학습데이터/테스트 데이터셋 분리는 8:2로 진행함.

Train 데이터 갯수: 42478

Test 데이터 갯수: 10620

**3-5. 사용 모델: BERT**

BERT (Pre-training of Bidirectional Transformers for Language Understanding)은 2018년에 구글에서 발표한 NLP Model 이다. 이 모델은 다양한 분야에서 뿐만이 아니라 특히 자연어 처리 분야에서 좋은 성능을 내고 있다. BERT는 일반적인 텍스트 데이터에 대해서 사전 학습을 하는 형식이다. General한 데이터에 대해 사전학습 (Pre-traine)한 후 원래의 목적으로 한 Dataset에 학습을 진행하여 다른 자연어 처리 모델들 보다 더욱 뛰어난 성능을 보인다.

일반 자연어 처리 모델: 분류 데이터 → LSTM/CNN등 머신러닝 모델 → 분류

BERT: 대량의 코퍼스 → BERT → 분류 데이터 → LSTM/CNN등 머신러닝 모델 → 분류

BERT는 총 3.3억 단어 (8억 단어의 BookCorpus 데이터와 25억 단어의 Wikipedia 데이터)의 거대한 코퍼스를 이용하여 학습한다고 한다. 이러한 사전 훈련 Embedding)을 통해 소비자 리뷰 데이터의 긍부정 지수를 얼만큼 정확하게 분류하는 지 알아보려고 한다.

**BERT에 맞게 전처리 후 Tokenizing 진행**

A screenshot of a computer

Description automatically generated with medium confidence

**BERT multilingual 내에 포함되어 있는 Tokenizer 사용**

Text

Description automatically generated

**Padding**

Text

Description automatically generated

**Attention Mask**

학습속도를 높이기 위해 실제 데이터가 있는 곳과 padding이 있는 곳을 attention에게 notify 한다고 한다.

Text

Description automatically generated

**\* GPU 할당**

해당 프로젝트는 Google Colab에서 개발되었기 때문에 Google Colab내에 있는 GPU를 할당 시켰음.

Text

Description automatically generated

사용 Optimizer는 transformers가 제공하는 AdamW / Epoch 수는 4로 지정

Text

Description automatically generated

**4. 학습 결과**

Text

Description automatically generated총 소요시간: 1h 43m 55s

정확도: 테스트셋 정확도: 82%

예시 (새롭게 생성 된 리뷰)

a. 부정

Graphical user interface, text

Description automatically generated

b. 긍정

Graphical user interface, text

Description automatically generated

**5. 마치며**

새로운 리뷰를 넣어 검증을 해 본 결과, BERT로 학습 된 모델이 긍부정 지수를 잘 분류하는 것을 확인할 수 있었음.

이 모델을 사용하여 추후 발생되는 리뷰 데이터의 긍부정 지수를 타 부서에게 전달하여 각 상황에 맞게 빠른 조치를 취할 수 있고 다양한 비지니스 인사이트 창출 예정.

또한 이 모델을 사용한 결과값을 대쉬보드화 하여 Sales/Marketing 팀에서도 확인할 수 있도록 개발 예정.

Github: <https://github.com/deardanny/ICT_AI_NLP/blob/main/review_sentimentanalysis.ipynb>

**5. 참고 자료**

BERT

<https://pangyo-datascientist.tistory.com/entry/BERT>

BERT로 영화리뷰 감성분류하기

<https://yesicakim.tistory.com/entry/E-19-BERT%EB%A1%9C-%EC%98%81%ED%99%94%EB%A6%AC%EB%B7%B0-%EA%B0%90%EC%84%B1%EB%B6%84%EB%A5%98%ED%95%98%EA%B8%B0>

딥러닝으로 동네생활 게시글 필터링하기

<https://medium.com/daangn/%EB%94%A5%EB%9F%AC%EB%8B%9D%EC%9C%BC%EB%A1%9C-%EB%8F%99%EB%84%A4%EC%83%9D%ED%99%9C-%EA%B2%8C%EC%8B%9C%EA%B8%80-%ED%95%84%ED%84%B0%EB%A7%81%ED%95%98%EA%B8%B0-263cfe4bc58d>

BERT로 네이버 영화 리뷰 데이터 분류하기

<http://yonghee.io/bert_binary_classification_naver/>

Kaggle: Sentiment Analysis using bert

<https://www.kaggle.com/code/adarshbiradar/sentiment-analysis-using-bert>

BERT를 활용한 속성기반 감성분석

<https://www.koreascience.or.kr/article/JAKO202009135419336.pdf>