Sentyle Project Poster

KOELECTRA를 활용하여 패션 쇼핑몰 리뷰 데이터 감성 분석 및 긍부정 예측

Authors
학번: 2021143039
이름: 윤주미

Affiliations
AI소프트웨어과

Introduction

패션은 우리 삶에서 더 이상 단순한 의복을 넘어 다양한 감성과 스타일을 표현하는 매체가 되었다. 이에 따라 온라인 패션 리뷰는 소비자들이 제품을 선택하고 유행에 민감하게 대응하는 데 큰 영향을 미치고 있다.

본 프로젝트는 소비자들이 온라인에서 더 나은 쇼핑을 할 수 있도록 도움을 주는 것을 목표로 하며, 이를 위해 AI Hub에서 제공하는 '쇼핑몰 리뷰 데이터'를 활용하여 리뷰의 긍부정을 예측하는 인공지능 모델을 개발하려고 한다.

KoELECTRA

프로젝트에서는 Hugging Face에 등록된 KoELECTRA-Small-v3 Discriminator의 토크나이저 및 PreTrained 모델을 사용했다.

KoELECTRA는 한국어 자연어 처리를 위해서 대규모 한국어 데이터를 사용하여 사전 훈련된 언어 모델이다.

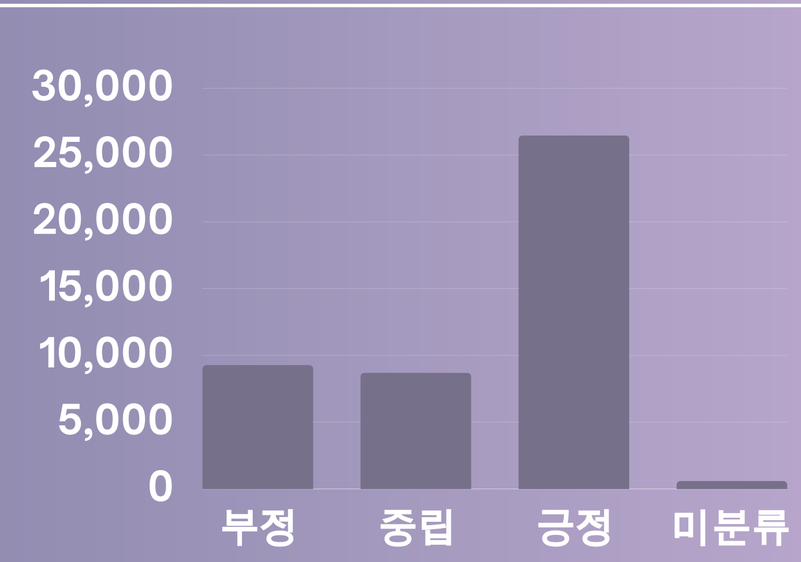
다양한 한국어 자연어 처리 작업에서 우수한 성능을 보이며, 오픈소스 프로젝트로 공개되어 한국어 자연어 처리의 정확도와 효율성을 높이는 데 활용된다.

Data

데이터는 2022년 쇼핑몰과 SNS 한글 리뷰 데이터로 구성되어 있고, 리뷰와 라벨링 데이터는 텍스트와 JSON 형태로 저장되어 있다.

이중에서 45,000건의 패션 분야 쇼핑몰 리뷰 데이터를 사용하여 프로젝트를 진행했다.

우측은 데이터의 긍부정 라벨 분포를 시각화하여 그래프로 나타내 보았다.



Data Preprocessing

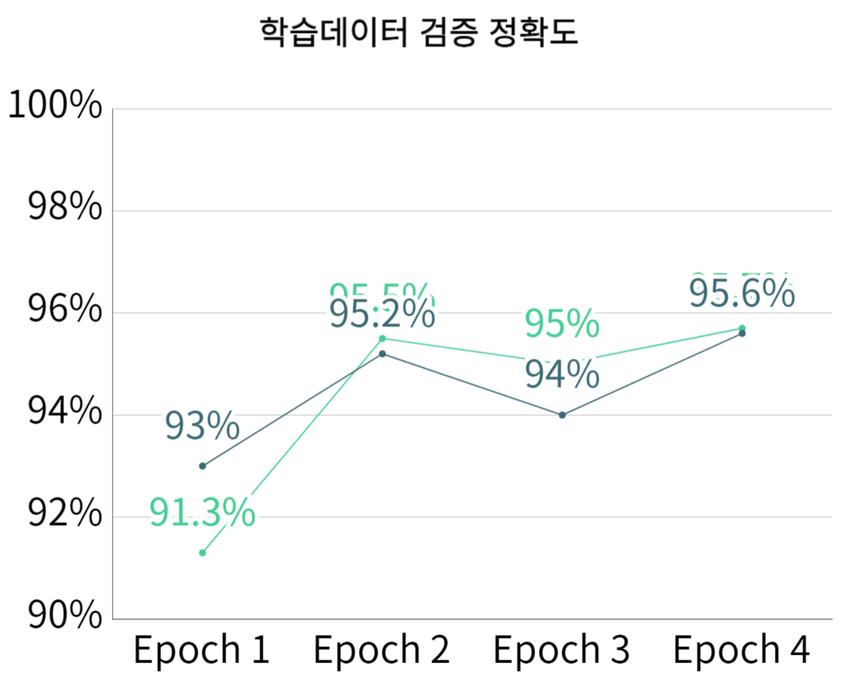
분류가 애매한 리뷰, 결측치와 중복값을 제거하였고, 15자 미만의 짧은 리뷰도 삭제하였다.

리뷰 중에 줄바꿈이 있는 경우 줄바꿈(\n)을 공백으로 대체하였고, label이 -1인 부정 리뷰를 label 0으로 변경하는 등의 전처리 작업을 수행하여 35,651건의 최종 데이터셋을 준비했다.

Analysis

모델의 성능을 높이고 새로운 데이터에 대한 예측 능력을 향상시키기 위해 최종 데이터셋을 학습과 검증 데이터셋으로 분리했다.

긍부정 리뷰를 각각 1,500개씩 뽑아 총 3000개의 데이터셋을 만들었고, 4대1 비율로 학습, 검증 데이터셋을 분리하였다.



Results

오차(loss) 값은 0.389 -> 0.04로 감소, 검증 정확도는 93% -> 95.6%로 증가하였다. 두 값은 학습이 진행됨에 따라 변화를 보인다.

각각의 변화는 모델이 학습 데이터로부터 효과적으로 학습되었다는 것과 리뷰의 긍부정을 잘 예측하고 있다는 것을 나타낸다.

전체 데이터에 모델을 적용한 결과 긍부정 예측 정확도가 0.96%로 높게 나왔다.