Final Project Outline

팀

A 팀

팀원

김서연, 정다솜, 박민서, 이준서

주제

뉴스기사의 주제별 분류와 감정 분석을 위한 머신러닝 모델 설계

주제 선정 배경 및 목적

오늘날처럼 복잡한 사회에서는 매일 수많은 뉴스 기사가 쏟아진다. 다양한 주제와 관점을 담은 기사들이 넘쳐나면서, 소비자들은 어떤 뉴스를 읽어야 할지 고민에 빠지기 쉽다. 특히 사건사고가 끊이지 않는 상황에서는 뉴스를 읽기 전에 미리 긍정적인 소식만 보고 싶은 순간도 있다. 이에 우리 조는 머신러닝 모델을 활용하여 뉴스의 카테고리를 자동으로 분류하고, 해당 뉴스가 긍정적인 내용인지 부정적인 내용인지 판단할 수 있도록 하여 소비자가 원하는 뉴스를 선별적으로 소비할 수 있도록 돕고자 한다.

데이터 출처 및 간단한 설명

1. 뉴스기사 주제별 분류에 사용할 데이터셋

- 다운로드 링크:
  https://www.kaggle.com/datasets/timilsinabimal/newsarticlecategories?utm_source=chatgpt.com
- 데이터 설명:
  Kaggle 의 'News Article Category Dataset'으로, HuffPost 라는 외국 언론사에서 발행된 뉴스기사 6877 개의 헤드라인과 본문을 바탕으로 총 14 개의 카테고리로 분류한다. 이때 분류되는 카테고리에는 'Arts and Culture', 'Business', 'Comedy' 등이 있다. 기존의 4 개의 혹은 41 개 카테고리로 분류된 다른 데이터셋과 비교했을 때, 14 개라는 숫자는 카테고리 수가 지나치게 적지도 많지도 않아 학습에도 적절하고 분석의 의미를 잃지도 않는다고 판단하여 이 데이터셋을 선택하게 되었다.

2. 뉴스기사 감정분석에 사용할 데이터셋

- 다운로드 링크:
  https://www.ims.uni-stuttgart.de/en/research/resources/corpora/goodnewseveryone/
- 데이터 설명:
  University of Stuggart 에서 연구를 진행하기 위해 사용한 'GoodNewsEveryone'이라는 데이터셋으로, 뉴스 헤드라인을 바탕으로 anger, sadness, disgust 등 다양한 감정으로 분류하였다. 데이터의 총 개수는 5000 개이다.

해결하고자 하는 문제 정의

뉴스 기사의 헤드라인 및 본문을 입력값으로 사용하고, 해당 뉴스 기사의 카테고리와 두드러진 감정을 출력값으로 예측하는 머신러닝 모델을 설계하고자 한다.