딥러닝 프로젝트 기획안

팀명	진실의 딥 - 4 조
주제명	AI 와 인간 리뷰 구분 모델

1. 역할분담

이름	역할
구경서	데이터 시각화, 프로젝트 발표 대본 작성 및 발표
김명진	데이터 시각화, 기획안 및 발표자료 검토
신민경	발표자료 준비, 모델 성능 평가
여혁수	모델 학습 및 하이퍼파라미터 튜닝, 결과 분석
최은서	데이터 전처리, 모델 구조 설계, 팀원들의 작업 조율

2. 주제 선정 배경

AI 의 리뷰 증가: 최근 AI 기반 리뷰 생성 기술이 발전하면서, 리뷰 플랫폼이나 전자상거래에서 AI 가 생성한 리뷰가 인간의 리뷰와 섞여 있을 가능성이 높아짐. 이러한 상황에서 허위 리뷰로 인해 소비자들이 잘못된 구매 결정을 내릴 가능성이 증가되고, 기업의 신뢰성에도 영향을 미침.

사회적 문제: 최근 온라인 리뷰에서 가짜 리뷰가 문제되고 있음. 특히, 일부 리뷰 알바들이 건당 수당을 받으며 제품에 대해 허위 리뷰를 작성하는 경우가 많음. 이로 인해 소비자들이 잘못된 정보를 바탕으로 구매 결정을 내리게 됨.

현행 규제 집행: 미국의 연방거래위원회는 온라인 허위 리뷰 문제의 심각성을 인식하고 강력한 규제 조치를 취하고 있음. 2024 년 8 월 '온라인 가짜 리뷰

금지법'을 최종 승인하고 허위 리뷰 1 건당 최대 5 만달러의 벌금을 부과함. 그러나 허위 리뷰를 식별하고 찾아내는 데 어려움을 겪고 있음.

딥러닝의 활용성: 딥러닝 모델은 대규모 텍스트 데이터에서 복잡한 패턴을 학습할 수 있어, AI 와 인간의 리뷰를 구분하는 데 강력한 도구가 될 수 있음.

3. 프로젝트 목표

- AI 와 인간이 쓴 리뷰를 자동으로 구분하는 딥러닝 모델 개발하고, 가짜 리뷰의 패턴을 파악하고 분석함.

- 한국어와 영어 텍스트 리뷰를 처리할 수 있는 모델을 각각 개발하여 다양한 언어 환경에 적용 가능한 시스템을 구축함.
- 현업에서 즉시 활용 가능한 수준의 모델을 도출함으로써 AI 와 인간 리뷰의 구분을 통한 소비자 보호와 기업 신뢰성 향상에 기여함.

4. 활용 데이터

1. Fake Reviews Dataset (ENG)

이 데이터셋은 고객의 리뷰 텍스트와 해당 리뷰가 AI 가 작성한 것인지, 인간이 작성한 것인지를 구분하는 레이블이 포함되어 있음.

<데이터셋 구성>

- category: 제품의 카테고리
- rating: 고객 평점
- label: AI('CG') 또는 인간('OR') 리뷰 여부
- text_: 리뷰 텍스트

2. DACON 음식점 리뷰와 게임 리뷰(KOR)

'AI vs Human 텍스트 판별 해커톤'에 사용된 데이터를 활용함. 4 개의 보기 중에 하나만 인간이 쓴 리뷰, 나머지는 AI 리뷰임. 이를 위해 한국어 처리에 적합한 전처리 과정이 필요함.

<데이터셋 구성>

- id : 샘플 고유 id
- sentence1 : 리뷰 텍스트 1
- sentence2 : 리뷰 텍스트 2
- sentence3 : 리뷰 텍스트 3
- sentence4 : 리뷰 텍스트 4
- label : 사람이 작성한 원본 리뷰 텍스트 번호; [1, 2, 3, 4] 중 하나

데이터 출처

Fake Reviews Dataset : https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset

음식점 리뷰와 게임 리뷰 : https://dacon.io/competitions/official/236178/overview/description

5. 분석 방안

- **분류 모델 전 불완전 데이터 확인 및 처리 모델 구현**

  AI 리뷰의 일부에서 중간에 잘린 리뷰 텍스트를 모델을 통해 추측하여 복원하는 작업도 포함됨. 이는 AI 리뷰의 특성을 반영하고, 텍스트가 불완전하게 제공될 때도 의미 있는 분석을 가능하게 할 수 있는 기술을 개발하는 데 중점을 둠. 또한, 뒷부분을 인공지능을 통해 생성한 데이터와 원본 데이터를 대상으로 동일한 모델을 적용함으로써 성능 비교가 가능해짐.

  베이스라인 모델로는 LSTM 을 사용할 것임. CNN + LSTM 구조로 CNN 은 문맥 특징을 추출하고, LSTM 으로 문장의 순차적 관계를 학습하는 방식의 모델을 고려할 예정.

- **분류 모델 선정 및 학습**

  자연어 처리(NLP) 모델로 LSTM(Long Short-Term Memory) 또는 BERT(Bidirectional Encoder Representations from Transformers)와 같은 딥러닝

모델을 사용하여 텍스트의 특징을 추출하고 AI 와 인간 리뷰를 구분함. MLP, CNN, LSTM 과 같이 강의 커리큘럼에 맞춘 기초적인 베이스라인 모델 구조를 다양하게 실험하고 성능을 비교, 최종적으로 모델 튜닝하면서 최적의 결과를 도출함.

- **성능 평가**

정확도, F1 스코어 등의 평가 지표를 통해 모델의 성능을 평가하고, 테스트 데이터로 최종 모델의 성능을 검증함.

- **하이퍼파라미터 최적화**

GridSearchCV 또는 Keras Tuner 등을 활용하여 모델의 하이퍼파라미터를 최적화하고, 과적합을 방지하기 위해 EarlyStopping 을 적용함.

6. 결 론

- **기대 효과**
  - 가짜 리뷰를 자동으로 탐지할 수 있는 시스템이 구축되면, 소비자와 기업 모두에게 신뢰를 제공할 수 있음.
  - 소비자에게 올바른 정보 제공을 통한 합리적 구매 결정을 지원할 수 있음.
  - 리뷰 데이터에서 텍스트 특징 분석(단어 사용 빈도, 문장 길이, 독창성 등)을 통해 인간 리뷰의 특징, 인사이트를 도출할 수 있으며, 리뷰 분류의 정확도를 높여 기업의 제품 품질 및 평판 관리를 개선할 수 있음.
- **미래 방향**
  - 프로젝트에서 개발된 모델을 실제 리뷰 플랫폼에 적용하여, 실시간으로 가짜 리뷰를 탐지하는 시스템을 구축할 수 있음.
  - 다양한 산업 분야(예: 전자상거래, 소셜 미디어)에서 활용될 수 있으며, 리뷰 분석 및 생성을 위한 중요한 기술로 자리 잡을 수 있음. 또한, 다양한 언어에서 적용 가능하여 글로벌 리뷰 시스템으로 확장될 수 있음.