ML 최종 보고서

과목: 기계학습의 이해

학번: 202400977

이름: 양수찬

1. 프로젝트 개요

1-1. 무엇을 만들었나

본 프로젝트는 'CS_Insight News Classifier'라는 이름의 웹 기반 도구/서비스입니다.

이 도구는 **머신러닝(ML) 모델**을 활용하여 전 세계의 실시간 RSS 뉴스 피드에서 수집한 헤드라인을 자동으로 분류하고, 특히 **컴퓨터 과학(CS) 전공자에게 유의미한 뉴스**를 선별하여 제공하는 서비스입니다.

주요 기능:

1. **실시간 뉴스 수집:** 설정된 복수의 RSS 피드(예: NYT Politics, BBC Business, Wired Technology 등)에서 최신 뉴스 헤드라인, 원본 카테고리 및 **원본 기사 링크**를 수집합니다.
2. **ML 기반 자동 분류:** 수집된 헤드라인을 학습된 분류 모델로 분석하여, 뉴스를 사전에 정의된 카테고리 중 하나로 예측합니다.
3. **CS_Insight 선별:** 예측된 카테고리 중 'CS_Insight'로 분류된 뉴스만 따로 모아 제공합니다.
4. **분석 및 통계:** 전체 뉴스 수, 'CS_Insight' 뉴스의 수, 그리고 'CS_Insight'로 분류된 뉴스들이 원래 어떤 매체의 카테고리(Politics, Economy, Technology 등)에서 왔는지에 대한 통계 정보를 제공합니다.
5. **텍스트 입력 분석:** 사용자가 직접 헤드라인을 입력하여 분류 모델의 예측 결과를 즉시 확인할 수 있는 기능도 제공합니다.

1-2. 어떤 문제를 해결했나

본 프로젝트의 'CS_Insight News Classifier'는 현대 디지털 환경에서 발생하는 가장 큰 문제인 정보 과부하와 정보 편향성 문제를 머신러닝 기반 필터링을 통해 해결합니다.

1. 정보 과부하 및 비효율적인 시간 소모 해결

매일 아침 신문사 별 뉴스를 볼 때 마다, it 와 관련된 뉴스들만 모아 보고 싶지만, it 와 관련됐어도 politics, economy 등 다양한 분야에 분포되어 있습니다. 따라서 관련 뉴스를 찾고 읽는데 많은 시간이 걸립니다.

- 방대한 데이터 스트림: RSS 피드와 같이 실시간으로 갱신되는 뉴스 스트림은 그 양이 엄청나서, 사람이 직접 모든 기사를 읽고 CS 관련 여부를 판단하는 것은 비현실적이며 시간 소모적입니다.
- ML 기반의 자동 선별: 본 도구는 ML 분류 모델을 활용하여 이 비효율성을 해소합니다. 설정된 복수의 뉴스 소스에서 수집된 수많은 헤드라인을 단 몇 초 만에 처리하고, 학습된 기준으로 'CS_Insight' 카테고리로 분류합니다.

- 극적인 시간 단축: 이 과정 덕분에 사용자는 관심 없는 일반 정치, 스포츠 뉴스와 같은 노이즈를 걸러내고, 최종적으로 'CS_Insight 뉴스 모아보기' 탭에서 핵심 기술 뉴스만을 모아서 확인할 수 있습니다. 이는 정보 탐색에 드는 시간을 극적으로 단축시키고, 사용자가 절약한 시간을 심층 분석이나 학습에 집중할 수 있도록 돕습니다.

2. 다양한 소스 통합을 통한 정보 편향성 극복

특정 신문사나 매체의 기술 섹션만으로는 전 세계적인 기술 동향의 입체적인 시각을 얻기 어렵고, 이는 정보 편향성으로 이어질 수 있습니다.

- 다중 RSS 피드 통합: 본 시스템은 단일 소스에 의존하지 않고, NYT(정치, 기술), BBC(경제, 사회, 스포츠), Wired(기술) 등 다양한 성격과 관점을 가진 전 세계 주요 언론사의 RSS 피드를 통합적으로 수집합니다.
- 종합적이고 균형 잡힌 시각 제공: 이처럼 광범위하고 다양한 매체로부터 데이터를 수집함으로써, 사용자는 한쪽으로 치우치지 않고 경제적, 정치적, 사회적 맥락에서 기술이 어떻게 논의되고 있는지를 종합적이고 균형 잡힌 시각으로 파악할 수 있습니다.
- 원천 카테고리 분석 제공: 나아가, 'CS_Insight 원본 카테고리 분석' 기능을 통해 기술 뉴스가 원래 어떤 분야(정치, 경제 등)에서 시작되었는지 통계적으로 보여주어, 기술이 사회 전반에 미치는 영향을 교차 분야적으로 이해하는 데 도움을 줍니다.

2. 진행 과정

2-1. 주제 선정 및 문제 정의

주제: ML 기반의 CS_Insight 뉴스 자동 분류 및 분석 도구 개발

문제 정의: 현대 사회의 정보 과부하 환경에서, 컴퓨터 과학(CS) 전공자가 전 세계 다양한 매체에서 쏟아지는 뉴스 스트림 속에서 핵심 기술 동향(AI, 데이터, 반도체, 사이버 보안 등)과 관련된 정보를 효율적이고 다양한 신문사로부터 편향되지 않은 방식으로 뉴스를 선별하고 모집하는 데 어려움을 겪는다는 문제를 해결하고자 했습니다.

- 목표: 일반 뉴스(정치, 경제, 사회 등)와 기술 뉴스(CS_Insight)를 정확하게 분류하는 ML 모델을 개발하고, 이를 실시간 뉴스 수집 기능과 결합하여 사용자에게 필요한 정보만을 집약적으로 제공하는 서비스를 구축합니다.

2-2. 데이터 수집 및 분석

- **학습 데이터:** 프로젝트 초기 단계에서 학습 데이터셋은 텍스트 분류를 위해 구성되었으며, 헤드라인 텍스트를 입력 특징으로, 카테고리(CS_Insight, Politics, Economy 등)를 레이블로 사용했습니다.
  이때, 기존 카테고리는 Politics, Society, Economy, Technology 4 개였으나, CS_KEYWORDS = ['ai', 'artificial intelligence', 'tech', 'technology', 'digital', 'data', 'cyber', 'crypto', 'semiconductor', 'nvidia', 'chip', 'google', 'microsoft', 'apple', 'meta', 'algorithm', 'software', 'cloud', 'robot', 'ev', 'electric vehicle'] 를 이용하여 4 개의 categories 에 분포돼 있는 cs 와 관련된 뉴스들을 cs_insight 카테고리에 재분류 했습니다.

2-3. 데이터 수집 및 전처리

항목	**내용**
데이터 소스	NYT, BBC, HuffPost, CNBC, Wired, The Verge 등 **13 개**의 글로벌 언론사 RSS 피드를 통합적으로 수집.
수집 항목	뉴스 **헤드라인 (headline)** 및 매체별 **원본 카테고리 (original_category)**
수집 결과	중복을 제거한 **총 409 개**의 고유한 뉴스 헤드라인 데이터 확보.
재분류 로직	헤드라인에 **20 개**의 CS 관련 키워드(예: 'AI', 'data', 'NVIDIA', 'robot' 등)가 포함될 경우, 원본 카테고리에 관계없이 최종 레이블을 **'CS_Insight'**로 재분류.

탐색적 데이터 분석 (EDA) 결과

수집된 데이터셋의 품질과 분류 모델 학습에 미치는 잠재적 편향 요소를 진단하기 위해 EDA 를 수행했습니다.

1. 데이터 무결성 및 구조 확인

확인 항목	결과	분석
중복 데이터	0 개	수집 단계에서 headline 기준 중복 제거 로직이 정상 작동하여, 모든 데이터가 고유함 (Clean Data)을 확인.
결측치	0 개	모든 컬럼(headline, original_category, category)에서 결측값이 없어(409 Non-Null Count), 추가적인 데이터 대체 작업 불필요.
데이터 구조	headline (입력 Feature), category (예측 Target)으로 구성.	

2. 클래스 분포 분석 (Class Distribution)

최종 레이블(category)의 분포를 확인하여 클래스 불균형 여부를 진단했습니다.

- **총 데이터 개수:** 409 개
- **클래스별 분포:**
    - **Majority Class:** Society (116 개)
    - **Minority Class (Target):** CS_Insight (57 개)

분석:

- 가장 데이터가 많은 클래스(Society, 116 개)와 가장 적은 클래스(CS_Insight, 57 개)의 비율은 약 **2:1** 수준으로, 학습에 심각한 지장을 줄 정도의 **극심한 클래스 불균형은 없는 것으로 판단**했습니다.
- 'CS_Insight' 카테고리는 특정 키워드 필터링을 통해 선별되었기 때문에 다른 일반 카테고리보다 데이터 수가 적게 나타나는 것은 예상된 결과입니다.

3. 텍스트 길이 분포 분석 (Text Length Analysis)

헤드라인의 글자 수와 단어 수 분포를 분석하여, 모델이 텍스트 '내용'이 아닌 '길이'로 분류하는 편향 가능성을 점검했습니다.

- **평균 헤드라인 길이:** 약 **62.27** 글자.

분석:

- **길이 균형:** 모든 5 개 카테고리의 **중앙값(Median)이 10~11 단어 부근**에 거의 일치했습니다.
- **결론:** 카테고리 간 텍스트 길이에 유의미한 차이가 없으므로, 향후 모델링 시 **길이 정보를 분류의 힌트로 삼는 편향 없이** 텍스트의 **실질적인 내용**에 집중하여 학습할 수 있는 고품질 환경이 조성되었음을 확인했습니다.

4. CS_Insight 키워드 정합성 검증

CS_Insight 카테고리로 분류된 기사들이 실제로 기술 도메인과 관련 있는지 확인하기 위해, 해당 카테고리 내 헤드라인의 상위 빈도 단어(Top 10 Keywords)를 분석했습니다.

분석:

- **주요 키워드: tech**, **data**, **google**, **robot** 등 CS 도메인과 직접적으로 관련된 단어들이 압도적인 빈도로 상위권을 차지했습니다.
- **정합성 결론:** 이는 **키워드 기반 재분류 알고리즘이 의도한 대로 정확하게 작동**하여, 방대한 뉴스 데이터 속에서 CS 전공자에게 필요한 정보를 효과적으로 필터링했음을 증명하며, 구축된 데이터셋이 **고품질의 Ground Truth** 로 활용하기에 적합하다는 결론을 내렸습니다.

2-4. ML 모델 학습 및 평가

이 단계에서는 Assignment 4 에서 구축된 뉴스 헤드라인 데이터셋을 활용하여 TF-IDF + MLP(Multi-layer Perceptron) 신경망 모델을 학습하고, 성능을 검증했습니다.

데이터 준비 및 분할

프로젝트의 안정적인 성능 측정 및 과적합 방지를 위해 데이터를 3 단계로 분할했습니다.

1. 레이블 인코딩: 최종 분류 타겟인 5 가지 카테고리(CS_Insight, Economy, Politics, Society, Technology)를 Label Encoder 를 사용하여 0 부터 4 까지의 정수 레이블로 변환했습니다.
2. 데이터 분할: 전체 409 개의 샘플을 클래스 비율을 유지하며(stratify=y) 분할했습니다.

데이터셋	비율	샘플 수	용도
Training Set	64%	245 개	모델 가중치 학습
Validation Set	16%	82 개	학습 중 성능 모니터링 및 조기 종료 판단
Test Set	20%	82 개	최종 성능 측정 (학습 완료 후 별도 평가)

모델 아키텍처 및 학습 설정

본 프로젝트는 텍스트 분류를 위해 파이프라인 (Pipeline) 구조를 사용했습니다.

- 특징 추출 (Input Layer): TfidfVectorizer 를 사용하여 텍스트 헤드라인을 수치 벡터로 변환했습니다. 영어 불용어를 제거하고, 최대 3,000 개의 특징(단어)만 사용하도록 설정했습니다.
- 분류기 (Classifier): MLPClassifier (다층 퍼셉트론 신경망)을 사용했습니다.
    - 은닉층 구조: 2 개의 은닉층, 각각 128 개와 64 개의 노드(Units)로 구성됨 (hidden_layer_sizes=(128, 64)).
    - 최적화/학습: adam, 초기 학습률 0.001 사용.
- 과적합 방지: 조기 종료 (early_stopping=True) 옵션을 활성화하여, 내부 검증 점수(Validation Score)가 10 Epoch 동안 개선되지 않을 경우 자동으로 학습을 멈추도록 설정했습니다.

학습 결과 및 평가

모델은 27 Epoch 만에 조기 종료되었습니다. 학습 과정에서 Loss 는 1.679 에서 0.0094 까지 감소했으며, test set 에 대한 최종 성능은 다음과 같습니다.

클래스 (레이블)	Precision	Recall	F1-Score	Support
CS_Insight (0)	0.89	0.73	0.80	11
Economy (1)	0.42	0.40	0.41	20
Politics (2)	0.54	0.58	0.56	12
Society (3)	0.56	0.62	0.62	23
Technology (4)	0.62	0.62	0.62	16

클래스 (레이블)	Precision	Recall	F1-Score	Support
Average (Weighted)	0.58	0.57	0.57	82

결론:

- 전반적 성능: 검증 셋에 대한 Accuracy 는 0.57 로, 5 개 클래스 중 무작위 추론(0.20)보다 훨씬 높은 성능을 보였으나, 개선의 여지가 있음을 확인했습니다.
- 핵심 카테고리 (CS_Insight): 본 프로젝트의 목표인 CS_Insight 카테고리에서는 F1-Score 0.80, Precision 0.89 를 기록하여, 이 도구가 CS 관련 뉴스를 비교적 높은 정확도로 선별하고 있음을 확인했습니다.

2-5. 최종 모델 저장

모델 학습 및 검증을 완료한 후, 추후 서비스 개발 및 배포를 위해 최종 모델을 외부 저장소에 안전하게 보관했습니다.

1. **객체 직렬화:** 성능 검증을 마친 최종 Pipeline 모델 객체와 레이블 인코더(Label Encoder) **news_classifier_model.pkl** 파일로 저장했습니다. 이 파일은 모델 추론에 필요한 모든 구성 요소(TF-IDF 벡터화기 포함)를 포함합니다.
2. **클라우드 저장소 배포:** 이 .pkl 파일을 **Google Drive** 에 업로드하여 퍼블릭 공유 링크를 생성했습니다.
3. **서비스 연결: Gradio 웹 서비스 코드(app.py)** 에는 Google Drive 의 공유 ID 를 포함한 **gdown** 다운로드 로직을 구현했습니다. 이로 인해 서비스가 Hugging Face Spaces 등 어떤 환경에서 실행되더라도, 별도의 수동 업로드 없이 **실행 시점에 자동으로 모델 파일을 다운로드하여 로드**할 수 있는 구조를 확립했습니다.

3. 모델을 서비스로 만든 구조

3-1. 모델을 서비스로 만든 구조

app.py 코드를 Hugging Face Spaces 에 배포함으로써, 프로젝트는 단순한 로컬 실행 코드가 아닌 **실제 사용자에게 접근 가능한 클라우드 기반 서비스**로 전환되었습니다.

- **배포 환경: Hugging Face Spaces** 를 활용했습니다. 이는 Gradio 애플리케이션을 위한 호스팅 환경을 제공하며, 사용자가 별도의 설치 과정 없이 웹 브라우저를 통해 언제든지 서비스에 접근하고 상호작용할 수 있도록 합니다.
- **시스템 구조**
    1. **지속적인 접근성:** Gradio 코드가 Hugging Face Spaces 의 서버에서 실행되면서, **퍼블릭 URL**(https://huggingface.co/spaces/Suchan544/cs_insight_classifier)을 통해 서비스가 24 시간 제공됩니다.
    2. **모델 로드:** 서비스가 시작될 때, 코드는 Google Drive 링크를 통해 news_classifier_model.pkl 을 다운로드하고 joblib.load()를 통해 모델을 로드합니다.
    3. **실시간 추론:** 사용자가 '뉴스 수집 및 분류 시작' 버튼을 클릭하면, Spaces 서버에서 requests 를 이용한 RSS 데이터 수집 및 **model.predict()** 추론이 실행되며, 그 결과를 Gradio 컴포넌트에 즉시 업데이트하여 웹 페이지에 표시합니다.
- **최종 사용자 경험:** 사용자들은 제공된 URL 에 접속하기만 하면 (PC 또는 모바일 환경에서) 바로 최신 뉴스를 분류하고 CS_Insight 결과를 확인할 수 있어, **효율적으로 cs 와 관련된 뉴스를 모아 볼 수 있습니다**.

4. 실제 사용 결과

(1)

CS_Insight 뉴스 모아보기

headline	original_category	link	Predicted_Catego
ICE Says It Has No Videos to Release of Chicago Deportation Operations	Politics	https://www.nytimes.com/2025/12/10/us/politics/ice-bodycam-video-chicago.html	CS_Insight
Judge in Oregon Blocks Arrest of Protesters For Noise	Politics	https://www.nytimes.com/2025/12/10/us/politics/trump-arrests-protesters-noise-ruling.html	CS_Insight
Oil Tanker U.S. Seized Off Venezuela Has Faked Its Location Before, Data Shows	Politics	https://www.nytimes.com/2025/12/10/us/politics/oil-tanker-venezuela-tracking-data.html	CS_Insight
History Colorado Center Rejects Painting, Citing Campaign Finance Law	Politics	https://www.nytimes.com/2025/12/10/arts/design/history-colorado-madalyn-drewno-free-speech.html	CS_Insight
Trump Administration Rules Threaten Nobel Prizes Won by Immigrants	Politics	https://www.nytimes.com/2025/12/10/science/nobel-prize-immigrants-science.html	CS_Insight
Shares in AI giant Oracle fall after revenue results ramp up bubble fears	Economy	https://www.bbc.com/news/articles/c9qe1e37411o?at_medium=RSS&at_campaign=rss	CS_Insight
Trump ban on wind energy permits 'unlawful', court rules	Economy	https://www.bbc.com/news/articles/cn7k6p6k5x5o?at_medium=RSS&at_campaign=rss	CS_Insight
Google unveils plans to try again with smart glasses in 2026	Economy	https://www.bbc.com/news/articles/cwyx83n00k6o?at_medium=RSS&at_campaign=rss	CS_Insight

(2)

CS_Insight 뉴스 모아보기

headline	original_category	link	Predicted_Category
Finance Law	Politics	madalyn-drewno-free-speech.html	CS_Insight
Shares in AI giant Oracle fall after revenue results ramp up bubble fears	Economy	https://www.bbc.com/news/articles/c9qe1e37411o?at_medium=RSS&at_campaign=rss	CS_Insight
Trump ban on wind energy permits 'unlawful', court rules	Economy	https://www.bbc.com/news/articles/cn7k6p6k5x5o?at_medium=RSS&at_campaign=rss	CS_Insight
Google unveils plans to try again with smart glasses in 2026	Economy	https://www.bbc.com/news/articles/cwyx83n00k6o?at_medium=RSS&at_campaign=rss	CS_Insight
'Carspreading' is on the rise - and not everyone is happy about it	Economy	https://www.bbc.com/news/articles/cy7vdvl2531o?at_medium=RSS&at_campaign=rss	CS_Insight
Why time is running out for Germany's green hydrogen industry	Economy	https://www.bbc.com/news/articles/cze60epnde0o?at_medium=RSS&at_campaign=rss	CS_Insight
The 'toughest crop': Can tech help cardamom farmers?	Economy	https://www.bbc.com/news/articles/cn0g2xxnrj3o?at_medium=RSS&at_campaign=rss	CS_Insight
Will AI mean better adverts or 'creepy slop'?	Economy	https://www.bbc.com/news/articles/ckg4y4z169go?at_medium=RSS&at_campaign=rss	CS_Insight

(3)

CS_Insight 뉴스 모아보기

headline	original_category	link	Predicted_Category
Trump Administration Rules Threaten Nobel Prizes Won by Immigrants	Technology	https://www.nytimes.com/2025/12/10/science/nobel-prize-immigrants-science.html	CS_Insight
Chip Company Plotted to Send Technology to China, Ex-C.E.O. Says	Technology	https://www.nytimes.com/2025/12/10/world/asia/dutch-nexperia-zhang-ceo.html	CS_Insight
Trump's Nvidia Chip Deal Reverses Decades of Technology Restrictions	Technology	https://www.nytimes.com/2025/12/09/us/politics/trump-nvidia-ai-chips-china.html	CS_Insight
App That Tracks ICE Raids Sues U.S., Saying Officials Pressured Apple to Remove It	Technology	https://www.nytimes.com/2025/12/08/business/apple-iceblock-lawsuit.html	CS_Insight
Meta Weighs Cuts to Its Metaverse Unit	Technology	https://www.nytimes.com/2025/12/04/technology/meta-cuts-metaverse-unit.html	CS_Insight
Silicon Valley Is All About the Hard Sell These Days	Technology	https://www.wired.com/story/sam-altman-jimmy-fallon-silicon-valley-hard-sell/	CS_Insight
America's Biggest Bitcoin Miners Are Pivoting to AI	Technology	https://www.wired.com/story/bitcoin-miners-pivot-ai-data-centers/	CS_Insight
Assisted dying bill: What is in proposed law?	Society	https://www.bbc.com/news/articles/cx2l7m6r55do?at_medium=RSS&at_campaign=rss	CS_Insight

(4)

CS_Insight 뉴스 모아보기

headline	original_category	link	Predicted_Category
Judge Orders Kilmar Abrego Garcia's Release From ICE Detention	Politics	https://www.nytimes.com/2025/12/11/us/politics/abrego-garcia-released.html	CS_Insight
ICE Says It Has No Videos to Release of Chicago Deportation Operations	Politics	https://www.nytimes.com/2025/12/10/us/politics/ice-bodycam-video-chicago.html	CS_Insight
Judge in Oregon Blocks Arrest of Protesters for Noise	Politics	https://www.nytimes.com/2025/12/10/us/politics/trump-arrests-protesters-noise-ruling.html	CS_Insight
Oil Tanker U.S. Seized Off Venezuela Has Faked Its Location Before, Data Shows	Politics	https://www.nytimes.com/2025/12/10/us/politics/oil-tanker-venezuela-tracking-data.html	CS_Insight
Shares in AI giant Oracle fall after revenue results ramp up bubble fears	Economy	https://www.bbc.com/news/articles/c9qe1e37411o?at_medium=RSS&at_campaign=rss	CS_Insight
'Architects of AI' named Time Magazine's Person of the Year	Economy	https://www.bbc.com/news/articles/c1y01mdm577o?at_medium=RSS&at_campaign=rss	CS_Insight
Trump ban on wind energy permits 'unlawful', court rules	Economy	https://www.bbc.com/news/articles/cn7k6p6k5x5o?at_medium=RSS&at_campaign=rss	CS_Insight
Google unveils plans to try again with smart glasses in 2026	Economy	https://www.bbc.com/news/articles/cwyx83n00k6o?at_medium=RSS&at_campaign=rss	CS_Insight

(5)

CS_Insight 뉴스 모아보기

headline	original_category	link	Predicted_Categor
App That Tracks ICE Raids Sues U.S., Saying Officials Pressured Apple to Remove It	Technology	https://www.nytimes.com/2025/12/08/business/apple-iceblock-lawsuit.html	CS_Insight
Crypto Magnate Do Kwon Sentenced to 15 Years in Prison	Technology	https://www.wired.com/story/do-kwon-terraform-sentenced-prison-crypto-fraud/	CS_Insight
Doxers Posing as Cops Are Tricking Big Tech Firms Into Sharing People's Private Data	Technology	https://www.wired.com/story/doxers-posing-as-cops-are-tricking-big-tech-firms-into-sharing-peoples-private-data/	CS_Insight
The Disney-OpenAI Deal Redefines the AI Copyright War	Technology	https://www.wired.com/story/disney-and-openais-deal-is-a-major-turning-point/	CS_Insight
Cursor Launches an AI Coding Tool For Designers	Technology	https://www.wired.com/story/cursor-launches-pro-design-tools-figma/	CS_Insight
AT&T's Connected Life Platform Is a Second Try in the Smart-Home Space	Technology	https://www.wired.com/story/att-connected-life-platform-launches-nationwide/	CS_Insight
Silicon Valley Is All About the Hard Sell These Days	Technology	https://www.wired.com/story/sam-altman-jimmy-fallon-silicon-valley-hard-sell/	CS_Insight
Assisted dying bill: What is in proposed law?	Society	https://www.bbc.com/news/articles/cx217m6r55do?at_medium=RSS&at_campaign=rss	CS_Insight

5. 배운 점 및 개선 방향

본 프로젝트를 통해 TF-IDF 벡터화와 MLP 신경망을 결합한 텍스트 분류 파이프라인을 구축하고, 최종적으로 Hugging Face Spaces 에 배포하여 **ML 모델의 서비스화**를 성공적으로 경험했습니다. 특히, 뉴스 헤드라인의 길이가 분류 편향을 유발하지 않는 균형 잡힌 데이터셋을 검증했으며, 'CS_Insight'와 같은 특화된 핵심 정보를 높은 정확도(F1-Score 0.80)로 추출하는 모델의 가치를 확인했습니다. 향후 개선 방향으로는 데이터 양을 누적 확보하고, 성능을 고도화하기 위해 **BERT 등 사전 학습된(Pre-trained) 모델로 전환**하는 것을 목표로 하며, 서비스의 활용성을 높이기 위해 **사용자 정의 키워드 필터링 기능**을 추가할 계획입니다.