한국어 문서 난이도 분석 및 어려운 단어 자동 번역 도구

이예송 202202605

1. 프로젝트 개요

본 프로젝트는 한국어 텍스트 내에서 사용자가 생소하게 느낄 수 있는 단어를 인공지능 모델로 예측하고, 해당 단어에 대한 스페인어 번역을 즉각적으로 제공하는 보조 도구 개발을 목적으로 합니다. 뉴스 기사나 학술적 문서를 읽을 때 모르는 단어가 나타날 때마다 사전을 검색해야 하는 번거로움을 해결함으로써, 독서의 흐름을 유지하고 한국어 텍스트에 대한 접근성을 높이고자 하였습니다.

2. 진행 과정

2.1. 주제 선정 및 문제 정의

본 프로젝트에서는 한국어 문서의 난이도를 분석하고, 사용자가 이해하기 어려울 것으로 예상되는 단어를 사전에 예측하여 자동으로 스페인어 번역을 제공하는 도구를 제작하였습니다.

스페인어권 환경에서 성장한 학습자의 경우, 한국 대학에서 제공되는 수업 자료를 읽는 과정에서 언어적 장벽을 느끼는 경우가 있습니다. 개념적으로는 이미 알고 있는 내용임에도 불구하고, 한국어로 접한 경험이 적은 어휘로 인해 단어 단위에서 이해가 단절되는 상황이 반복됩니다.

이러한 학습 환경에서는 문서를 전체 번역하는 방식보다, 이해에 문제가 되는 핵심 어휘만을 선별하여 학습하는 방식이 더 효과적일 수 있다고 판단하였습니다. 그러나 실제 학습 과정에서 매번 사전을 찾아보는 방식은 학습 흐름을 끊고 효율을 저하시킨다는 한계가 있습니다.

이에 본 프로젝트에서는 사용자의 어휘 인지 상태를 반영한 AI 모델이 문서 내에서 어려운 단어를 사전에 식별하고, 해당 단어에 대한 스페인어 번역을 제공함으로써 보다 효율적인 어휘 학습을 지원하고자 하였습니다.

2.2. 데이터 수집 및 분석

본 프로젝트의 핵심은 사용자 개인의 어휘 인지 수준을 모델에 반영하는 데 있으므로, 기존 공개 데이터셋을 사용하는 대신 데이터를 직접 수집하고 수작업으로 레이블링을 진행하였습니다. 데이터 수집은 총 3차에 걸쳐 이루어졌으며 모든 단어는 엑셀 파일로 정리한 뒤, 사용자가 알고 있는 단어는 1, 모르는 단어는 0으로 분류하였습니다.

초기 데이터셋을 활용한 실험 결과, 일부 일상적인 단어까지 모름으로 분류되는

현상이 확인되었습니다. 이는 모름 클래스 데이터의 비중이 충분하지 않아 예측 기준이 안정적으로 형성되지 못한 것으로 판단하였습니다. 이에 따라 데이터 분포의 분균형을 완화하고자 데이터 수집을 단계적으로 확장하여 모름 클래스의 비중을 보완하였습니다.

데이터 종류	단어 수	레이블	모르는 단어 비율
1차			
경제뉴스	268	0 = 27, 1 = 241	10%
일반뉴스	205	0 = 24, 1 = 181	11.7%
중남미경제 (수업)	515	0 = 97, 1 = 418	18.8%
문예사조의 이해 (수업)	399	0 = 57, 1 = 342	14.2%
합계	1387	0 = 205, 1 = 1182	14.8%
2차			
토픽 6급 어휘 목록	2327	0 = 791, 1 = 1536	34%
3차			
한자어 기단 단어 모음	307	0 = 190, 1 = 117	61.9%

2.3. ML 모델 학습 및 평가

2.3.1. 모델 선정 및 학습 환경

단어의 난이도를 파악하기 위해, 한국어 특화 사전 학습 모델인 KLUE/BERT-base를 기반 모델로 선정하였습니다. 학습은 GPU 가속이 가능한 Google Colab 환경에서 진행되었으며, Hugging Face Trainer API를 활용하여 효율적인 학습 파이프라인을 구축하였습니다. 모델의 출력층은 단어의 난이도에 따라 모름(0)과 앎(1)의 두 가지 클래스를 분류하도록 재구성하였습니다.

2.3.2. 데이터셋 구축

모델 학습을 위해 준비된 데이터셋은 총 4,021개의 단어로 구성되어 학습의 객관성을 확보하고자 전체 데이터를 8:1:1(학습:검증:시험)의 비율로 분할하였습니다. 특히 단어의

분포가 편향되지 않도록 Stratified Split 기법을 적용하여 각 세트 내의 라벨 비율을 일정하게 유지하였습니다. 각 단어는 BERT 토코나이저를 통해 최대 64토큰 길이의 텐서 데이터로 변환되어 모델에 입력되었습니다.

2.3.3. 모델 학습 및 최적화

[1206/1206 03:08, Epoch 3/3]

Epoch	Training Loss	Validation Loss
1	No log	0.459919
2	0.525600	0.476520
3	0.370400	0.628442

모델 학습은 총 3회의 에포크 동안 진행되었으며 Epoch 1에서 검증 손실 0.459로 시작하여 최종 Epoch에서 학습 손실이 0.370까지 안정적으로 하락하였습니다. 학습 과정 중 검증 손실이 가장 낮았던 시점의 가중치를 최종 모델로 저장하여 모델의 일반화 성능을 높이고 과적합을 방지하였습니다.

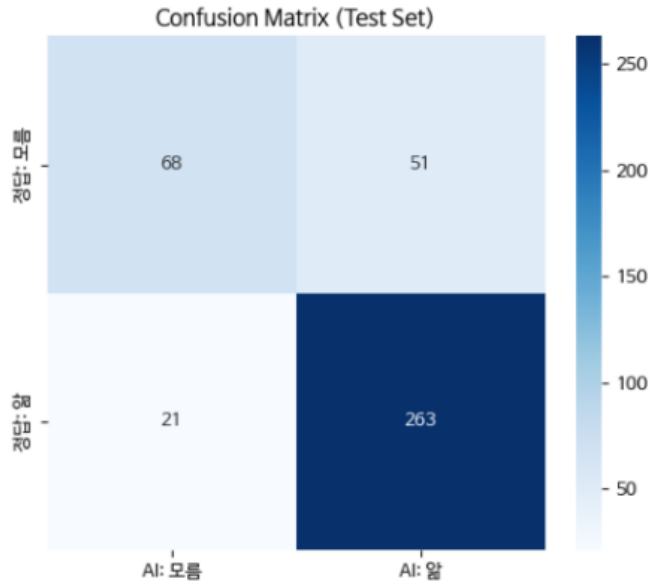
2.3.4. 성능 평가 및 결과 분석

```
======== 💯 최종 성적표 ========
정확도(Accuracy): 82.13%
------------------------------
상세 리포트:
              precision    recall  f1-score   support

      모름(0)       0.76      0.57      0.65       119
       앎(1)       0.84      0.93      0.88       284

    accuracy                           0.82       403
   macro avg       0.80      0.75      0.77       403
weighted avg       0.82      0.82      0.81       403
```

학습에 참여하지 않은 별도의 시험 데이터 403건을 대상으로 최종 성능을 측정한 결과, 약 82.13%의 정확도를 확보하였습니다. 아는 단어에 대한 재현율은 0.93으로 매우 우수하여 실사용 시 쉬운 단어는 대부분 걸러졌습니다. 모르는 단어 세션 역시 0.76의 준수한 정밀도록 기록하였습니다.



혼동 행렬 확인 결과, 실제 모름 데이터를 앎으로 오판한 사례가 51건 존재하였으나, 전체적인 분류 성능은 서비스 목적을 충족하는 수준으로 판단되었습니다.

가중 평균 F1-Score 0.81의 안정적인 성능을 바탕으로, 실제 서비스에서 어려운 단어를 선별하고 번역 정보를 제공하는 핵심 엔진으로서 충분한 신뢰도를 확보하였습니다.

전체 정확도는 82% 수준이었으나, 본 프로젝트의 목적이 '어려운 단어를 놓치지 않고 탐지하는 것'임을 고려하면 Accuracy보다 각 클래스의 Recall이 더 중요한 지표라고 판단하였습니다.

모름(0) 클래스의 재현율이 0.57로 나타난 것은 여전히 일부 어려운 단어가 탐지되지 않는다는 의미이며, 추후 모델 개선 시 우선적으로 보완해야 할 부분입니다. 반면 앎(1' 클래스의 재현율이 0.93으로 높게 나타난 것은 쉬운 단어가 데이터셋에서 상대적으로 많이 등장하고 BERT가 빈도 기반 패턴을 잘 학습했기 때문으로 보입니다.

3. 모델을 서비스로 만든 구조

3.1. 학습한 모델의 서비스화 과정 (web application)

모델을 실제 서비스로 제공하기 위해 로컬 개발 환경에서 전체 시스템 구조를 설계하고 구축하였습니다. 학습된 모델 가중치, 토크나이저 설정 파일, 데이터셋, 그리고 웹 UI 구현을 위한 app.py 코드를 하나의 프로젝트 폴더에 정리한 후 VS Code로 불러왔습니다.

터미널에서 streamlit run app.py 명령을 통해 애플리케이션을 실행하였으며 웹 페이지 형태의 사용자 인터페이스가 생성되어 입력 텍스트에 대해 단어 단위 분석 및 번역 결과를 실시간으로 확인할 수 있었습니다. 이를 통해 모델 추론 기능과 UI의 연동이 정상적으로 이루어지는지 검증하였습니다.

로컬 테스트 이후, 이를 로컬환경 뿐만 아니라, 어느 환경에서든 접속을 가능하게 하기 위해 프로젝트를 GitHub 저장소와 연동하였습니다. 다만 BERT 기반 모델의 용량이 커서 직접 업로드가 불가능했기 때문에, 모델 파일은 HuggingFace Hub에 업로드하였고 웹 서비스에서는 HuggingFace에서 모델을 직접 불러오는 방식으로 해결하였습니다.

최종적으로 Streamlit Cloud와 GitHub 저장소를 연결하여 베포함으로써, 로컬 환경이 아닌 웹 브라우저에서도 누구나 접근 가능한 형태의 서비스로 확장할 수 있었습니다.

3.2. 시스템 구조

3.2.1. Streamlit UI 레이어

사용자가 웹 화면에 텍스트를 입력합니다

3.2.2. 모델 로딩.추론 레이어

HuggingFace Hub에서 fine-tuned모델과 tokenizer를 로드 하여 입력된 텍스트를 단어 단위로 분리합니다. 그 후, 각 단어를 앎(1) 또는 모름(0) 클래스로 분류한 뒤 모름

단어는 스페인어 번역을 수행합니다.

3.2.3. 출력 레이어

모름 단어를 노란색으로 강조 표시하며 번역 결과를 출력합니다.

4. 실제 사용 결과

5가지 뉴스 기사로 실제 사용을 진행하였으며 경제 뉴스 2개, 정치 뉴스 2개, IT/과학 뉴스 1개를 준비하였습니다.

4.1. 뉴스

4.1.1. 경제 뉴스

해외 파생상품 투자하려면...모의거래 3시간 이수해야 가능

https://n.news.naver.com/mnews/article/016/0002571717

앞으로 개인투자자는 사전교육과 모의거래를 이수해야 해외 파생상품을 거래할 수 있다.

금융감독원은 15일부터 해외 파생상품을 처음 거래하는 일반 개인투자자에게 사전교육(1시간 이상)과 모의거래(3시간 이상) 이수를 의무화한다고 14일 밝혔다.

해외 레버리지 상장지수상품(ETP)을 처음 거래하려는 개인도 사전교육(1시간)을 받아야 한다.

사전 교육은 동영상으로 진행되며, 금융투자협회 학습 시스템을 통해 수강할 수 있다.

금감원은 해외 파생상품이 원금 초과 손실이 발생할 수 있는 고위험 상품이라고 강조했다.

예상치 못한 환율 변동으로 손실이 커질 수 있으며, 시세 급변 시 투자자 동의 없이 반대매매가 실행될 수 있다는 점도 유의해야 한다.

금감원에 따르면 2020년부터 올해 10월까지 개인투자자들은 해외 파생상품에서 연평균 약 4490억원의 손실을 봤다. 시장 등락과 무관하게 손실이 반복된 점이 특징적이다.

금감원은 "미국 증시(나스닥)가 큰 폭으로 하락한 2022년(-33.1%)뿐 아니라, 상승한 2020년(+43.6%) 및 2023년(+43.4%)에도 개인투자자는 큰 손실을 봤다"고 설명했다.

해외 파생상품 거래는 개인투자자 비중이 82.5%로 대부분을 차지하며, 개인 거래는 변동성 장세에서 활발해지는 경향도 나타났다.

국내 투자자가 보유한 해외 레버리지 ETP 규모도 2020년 이후 매년 급증하고 있다. 지난 10월 말 기준 19조4000억원으로 역대 최대치를 기록했다.

Word	Spanish Translation	Confidenc
개인투자자	inversor individual	66.25%
모의거래	comercio simulado	84.81%
이수해야	Debe completar	59.39%
파생상품	derivados	82.72%
레버리지	aprovechar	74.40%
상장지수상품(ETP)	Productos negociados en bolsa (ETP)	72.01%
금융투자협회	Asociación de inversión financiera	73.10%
수강할	para tomar el curso	62.31%
원금	principal	64.42%
고위험	alto riesgo	68.90%
반대매매	venta inversa	78.28%
연평균	Promedio anual	63.48%
등락	sube y baja	80.11%
장세	Precio de mercado	82.40%
활발해지	Ser activo	74.88%
ETP	ETP	64.58%

"구글의 무서운 추격"...챗GPT 독주 끝내고 '2강 체제' 굳히나

https://n.news.naver.com/mnews/article/055/0001316075

국내 AI 챗봇 시장이 최근 3주 동안 거센 지각변동을 겪고 있습니다.

오픈AI의 챗GPT가 여전히 독주하는 가운데 구글의 새 모델 '제미나이3'가 맹추격하고 있지만, 퍼플렉시티 이용자는 급격히 줄어드는 모양새입니다.

데이터 테크 기업 아이지에이웍스의 모바일인덱스는 구글이 제미나이3를 출시한 지난달 17일부터 3주간 국내 AI 챗봇 시장을 분석했습니다.

국내 선두를 달리는 챗GPT의 주간 활성 이용자는 지난달 17일부터 23일까지 869만 3천560명으로 집계됐습니다.

이후 2주 동안에도 각각 880만 475명과 875만 4천798명을 기록하며 압도적인 규모를 유지했습니다.

다만 챗GPT 신규 설치 건수는 첫째 주 20만 2천303건에서 셋째 주 19만 1천339건으로 소폭 하락한 것으로 나타났습니다.

반면 제미나이는 지난달 17일 이후 빠르게 이용자를 늘리고 있습니다.

제미나이의 주간 활성 이용자는 첫째 주 1만 6천196명에서 시작해 그 다음 주 2만 2천928명으로 급증했습니다.

신규 설치 건수 역시 첫째 주 5만 967건이었지만 그 다음 주에는 11만 1천115건으로 두 배 이상 뛰었습니다.

이는 새로운 기능 공개와 공격적인 홍보 효과로 초기 유입자가 늘어난 결과로 풀이됩니다.

이에 비해 AI 검색 분야 주요 모델인 퍼플렉시티는 이용자와 신규 설치 모두 뚜렷한 감소세를 보였습니다.

퍼플렉시티 이용자는 첫째 주 45만 5천659명에서 셋째 주에는 43만 6천480명까지 떨어졌습니다.

같은 기간 신규 설치 건수도 1만 6천908건에서 1만 2천134건으로 크게 줄어든 것으로 확인됐습니다.

전문가들은 챗GPT 독주 체제가 저물고 장기적으로 제미나이와 함께하는 'AI 2강' 체제가 도래할 것으로 전망했습니다.

Word	Spanish Translation	Confidence
거센	rígido	60.04%
지각변동	cambio tectónico	55.72%
오픈AI의	Abrir IA	83.78%
독주하	solo	83.55%
퍼플렉시티	Lexidad morada	61.00%
줄어드	disminuido	68.15%
아이지에이웍스의	Obras IGA	79.62%
모바일인덱스	índice móvil	61.25%
출시한	liberado	51.73%
소폭	levemente	57.78%
제미나이	Géminis	82.63%
제미나이의	Géminis	81.90%
유입자	afluencia de gente	60.41%
풀이됩니다	esta solucionado	54.54%
감소세	rechazar	75.66%
독주	solo	71.04%
도래할	venir	79.23%

4.1.2. 정치 뉴스:

유시민 "민주당, 뭐 하는지 모르겠다…지금 굉장히 위험" 경고

https://n.news.naver.com/mnews/article/661/0000067125

유시민 전 노무현재단 이사장이 최근 더불어민주당의 행보를 두고 "c최근 몇 달 동안 뭘 하는지 모르겠다"며 "지금 민주당은 굉장히 위험하다"고 직격했습니다.

유 전 이사장은 어제(13일) 열린 노무현재단 후원회원의 날 행사에 참석해 이 같이 밝혔습니다.

그는 "(민주당은) 왜 권한이 있는데 뭘 안 하고 말만 하느냐"며 "백날 토론만 하고 있지 말고 내란전담재판부를 만드는 법이든 뭐든 입법안을 내서 자기들이 해야 한다"고 지적했습니다.

이어 "대통령실과 의견이 맞네, 안 맞네 왜 그런 소리를 하느냐"며 "이재명 대통령이 '그런 걸 왜 당에서 마음대로 하느냐'고 할 분도 아니고, 의견이 다르더라도 '의원들이 당원들 뜻을 모아서 했다면 내가 받아들여야지'할 분"이라고 말했습니다.

내년 6월 지방선거를 두고는 "여당은 여당답게 시민들의 삶을 개선하는 데 초점을 맞추면 된다"고 조언했습니다.

조국혁신당과 관련해서는 "민주당이 지난 몇 달처럼 흐리멍덩한 태도를 취하면 조국혁신당에 기회가 생긴다"며 "조국혁신당은 '매운맛 민주당'"이라고 평가했습니다. 특히 "이대로 가면 호남에서(민주당이 조국혁신당과 맞붙을 경우) 위험하다"고 강조했습니다.

이재명 대통령에 대해서는 "우선 사람이 똑똑하다"며 "이거(대통령직)를 정말 오래 하고 싶어 했던 분인데, 하고 싶었던 분이 똑똑하기까지 하다"고 말했습니다. 이어 "지난 6개월 동안 굉장히 어려운 고비를 상당히 잘 넘겼다"고 평가했습니다.

Word	Spanish Translation	Confidence
유시민	Yoo Si Min	67.21%
노무현재단	Fundación Roh Moo Hyun	68.80%
행보	acción	75.31%
c최근	cReciente	57.62%
직격했습니다	fue un golpe directo	58.46%
후원회원의	de miembros patrocinadores	64.48%
내란전담재판부	Tribunal de insurrección	80.98%
입법안	legislación	71.60%
마음대	lo que quieras	60.99%
당원들	miembros del partido	59.35%
여당	partido gobernante	63.86%
조국혁신당	Partido de la Innovación de la Patria	68.21%
맞붙	Cara a cara	61.95%

조국 "'계엄 사과' 25 명 국힘 의원들, 뛰쳐나와 창당하라"

https://n.news.naver.com/mnews/article/666/0000090559

조국 조국혁신당 대표가 '12·3 비상계엄'에 대해 사과한 국민의힘 의원들 25명에게 국민의힘 탈당과 신당 창당을 제안했다.

조 대표는 14일 자신의 페이스북에 "윤석열 국회 탄핵 1주년인 오늘, 저는 25명의 국회의원에게 정중히 제안한다. 극우본당에서 뛰쳐나와 새로운 보수정당을 창당하라"고 밝혔다.

그는 "국회가 '내란 수괴 대통령 윤석열 탄핵소추안'을 가결한 지 1년이다. 그날의 탄핵안 통과는 응원봉 시민의 함성에 국회가 응답한 결과"라며 "4·19부터 5·18, 6·10, 촛불혁명, 응원봉 혁명까지 광복 후 약 80년 동안 우리 민주주의는 어떤 불의도 용납하지 않았다"고 말했다.

이어 "12월 14일은 민주주의 강국 대한민국의 힘을 전 세계에 보여준 역사적인 날"이라며 "그러나 1년이 지난 지금도 내란 수괴 윤석열은 내란의 술독에 빠져 내란의 정당성을 강변하고 있다. 내란 잔당 국민의힘은 내란의 숙취에 깨어나지 않고 극우본당으로 활개치고 있다"고 주장했다.

조 대표는 "그나마 지난 3일 국민의힘 25명의 의원이 용기를 냈다. 불법 계엄에 사과했고, 윤석열과의 단절을 선언했다"면서 "윤석열 국회 탄핵 1주년인 오늘, 저는 25명의 국회의원에게 정중히 제안한다. 극우본당에서 뛰쳐나와 새로운 보수정당을 창당하라"고 촉구했다.

그러면서 "국민에게 총을 겨눈 정당에서 도대체 어떤 정치를 하겠다는 것인가"라며 "'당 안에서 혁신하겠다'는 말은 '국회의원직만은 유지하겠다'는 비겁한 자기변명에 불과하다"고 목소리를 높였다.

이어 "김상욱 의원의 건강한 보수 정치의 용기를 본받기를 바란다. 25명이니 원내교섭단체도 가능하다"며 "탈당하고 새롭게 시작하라"고 재차 강조했다.

한편 이날 혁신당은 윤 전 대통령과의 단절 선언에 동참하지 않은 국민의힘 의원 82명의 지역구 사무실 앞에서 피켓 시위를 벌이고 항의 서한도 전달할 계획이다.

Word	Spanish Translation	Confidenc
조국	patria	55.27%
조국혁신당	Partido de la Innovación de la Patria	68.21%
탈당	atornillar	65.32%
신당	santuario	70.64%
창당	establecimiento	65.96%
"윤석열	"Yoon Seok Yeol	76.23%
탄핵	acusación	52.22%
극우본당	extrema derecha	63.83%
'내란	Guerra civil	72.37%
수괴	masa de agua	76.42%
윤석열	Yoon Seok Yeol	87.19%
탄핵소추안'	Propuesta de juicio político	56.23%
가결한	aprobado	77.43%
탄핵안	Moción de acusación	61.62%
시민의	civil	51.00%
촛불혁명	revolución a la luz de las velas	52.45%
광복	liberación	68.87%
불의도	La injusticia también	72.39%
강국	central eléctrica	79.41%
내란의	guerra civil	76.83%
강변하고	Por el río	83.60%
잔당	restos	76.41%
계엄	ley marcial	82.67%
윤석열과의	Con Yoon Seok Yeol	79.90%
'국회의원직만	Sólo como miembro de la Asamblea Nacional	71.57%
자기변명	autoexcusa	71.45%
"김상욱	"Kim Sang-wook	79.09%
원내교섭단체도	Grupo de negociación interno	67.57%
혁신당	fiesta de innovación	71.98%
서한도	Seohando	91.06%

4.1.3. IT/과학

LG유플러스, '구글 AI 프로' 제휴 상품 출시…최대 반값 할인 혜택

https://n.news.naver.com/mnews/article/009/0005605695

LG유플러스는 '구글 AI 프로(Google AI Pro)' 제휴 상품을 출시했다고 14일 밝혔다.

LG유플러스 이용자 대상으로 오는 30일까지 '구글 AI 프로' 모바일 부가서비스 가입 시 50% 할인된 가격에 구글의 AI 서비스 제미나이 3와 클라우드 저장공간 2TB를 이용할 수 있다.

'구글 AI 프로'는 △제미나이 3 △특화 이미지 생성 모델 나노바나나 프로 △동영상 제작 도구 플로우&위스크(Flow&Whisk) △전문적 수준 보고서 작성 기능 딥리서치(Deep Research) △리서치 및 학습 도구 노트북LM △2TB 클라우드 저장 공간 등 구글의 핵심 AI 기능과 모델을 이용할 수 있는 월정액 상품이다.

'구글 AI 프로'는 모바일 전용 요금제로도 이용 가능하다. 너겟 요금제 중 너겟65(데이터 80GB, 월 6만5000원)와 너겟69(데이터 무제한, 월 6만900원)에 가입한 이용자는 추가 비용 없이 사용할 수 있다. 데이터 무제한 5G 요금제 고객 역시 내년 1월부터 프미리어 서비스로 '구글 AI 프로'를 선택할 수 있다.

LG유플러스 관계자는 "앞으로도 구글과 지속적으로 협업해 고객 편의성을 확대할 다양한 서비스를 선보이겠다"고 말했다.

Word	Spanish Translation	Confidence
부가서비스	Servicios adicionales	61.10%
제미나	Gémina	81.08%
△제미나	△Jemina	70.81%
△특화	△Especialización	75.33%
나노바나나	nanoplátano	80.50%
플로우&위스크(Flow&Whisk	Fluir y batir	79.80%
△전문적	△Profesional	76.47%
딥리서치(Deep	Investigación profunda	71.48%
△리서치	△Investigación	83.39%
월정액	cuota mensual	52.97%
프미리어	Primer ministro	76.04%

4.2. 사용자 관찰 기반 긍정적 효과

본 서비스는 학습된 모델을 기반으로 다양한 주제의 뉴스 기사를 입력하여 실제 사용성을 검증하였으며 테스트 결과, 모델은 한자어 기반의 용어, 금융·정치·법률 분야의 전문 용어 등 사용자가 평소 이해하기 어려워하던 단어들을 비교적 정확하게 식별해냈습니다.

특히 모의거래, 레버리지, 등락, 장세, 지각변동, 유입자, 도래하다, 직격하다, 당원, 입법안, 강변하다, 잔당과 같이 일상적으로 자주 접하지 않는 전문 용어나 한자어 기반 표현들을 '모름(0)'으로 분류한 결과는 실제 사용 경험과도 높은 일치도를 보였습니다.

또한 행보, 가결처럼 자주 접하지만 정확한 의미가 헷갈려 확실히 뜻을 알고 싶은 단어들 역시 모름으로 감지되었습니다. 이는 모델이 단순히 생소한 단어뿐 아니라 의미적 판단이 필요한 어휘까지 선별해내는 기능적 장점을 보여주는 결과였습니다.

4.3. 한계 및 문제점

그러나 테스트 과정에서는 다음과 같은 한계점과 오류도 명확하게 확인되었습니다.

4.3.1. 지나치게 쉬운 단어를 '모름'으로 분류하는 경우

'시민', '내란', '마음대로', '풀이되다', '출시하다', '수강'처럼 기본 어휘로 분류되는 단어들도 모름으로 판단하는 경우가 있었습니다. 이는 학습 데이터에서 단어 난이도 기준이 완전히 정교하지 않았거나, 문맥 없이 단일 단어만 입력하는 구조적 한계 때문인 것으로 판단하였습니다.

4.3.2. 고유명사를 모름으로 분류하는 문제

'오픈AI', '제미나이', '나노바나나', '아이지에이웍스', 등 학술적 의미가 없는 고유명사(사람 이름, 회사명)를 어려운 단어로 탐지하였습니다. 이는 모델이 단어의 성격을 구분하지 못하는 구조적 한계에서 발생한 것으로 보입니다.

5. 배운 점 및 개선 방향

이번 프로젝트를 통해 문제를 기술적으로 정의하여, 데이터 구조로 구체화하는 과정의 중요성을 경험하였습니다. 모델 학습부터 평가, 그리고 실제 서비스 형태로의 배포까지 전체 흐름을 처음부터 끝까지 직접 수행하면서, AI 시스템이 어떤 단계들을 거쳐 작동하는지 실질적으로 이해할 수 있었습니다.

특히 VS Code 환경에서의 개발, 모델 가중치 관리, GitHub 및 HuggingFace를 활용한 배포 과정 등을 모두 거치며, 기술 도구를 스스로 선택하고 조합할 수 있는 방법을 배울 수 있었습니다.

또한, 그동안 반복적으로 겪어왔던 언어 장벽 문제도 AI를 활용하면 충분히 자동화하거나 개선할 수 있다는 점을 확인하였고 앞으로도 유사한 학습·언어적 어려움이 발생한다면 이번 프로젝트와 같은 방식으로 문제를 해결해볼 수 있겠다는 생각이 들었습니다.

그와 동시에 이번 프로젝트에서는 명확한 한계점도 확인되었습니다. 최종 정확도가 80%대 초반에 머물면서, 쉬운 단어가 모름으로 잘못 분류되거나, 반대로 특정 고유명사가 어려운 단어로 처리되는 경우가 여전히 존재했습니다. 이러한 부분은 데이터의 편향, 전처리 미흡, 문맥 정보 부족 등이 영향을 준 것으로 보입니다.

무엇보다 스페인어 번역 역시 단어 단위 번역만 가능해, 문맥이나 문장 구조에 따라 자연스러운 의미 전달이 어려운 경우도 확인되었습니다. 이 부분은 문장 단위 번역 모델과의 연동 또는 후처리 로직을 추가함으로써 개선할 수 있을 것으로 보입니다.

종합적으로 이번 프로젝트는 AI 모델 개발과 배포 전 과정을 직접 다뤄보는 경험이었고, 문제 정의·데이터 구성·서비스 구현까지 전체 흐름을 실제로 이해하는 데 큰 도움이 되었습니다. 동시에 모델 성능을 높이기 위한 구체적인 개선 방향도 확인할 수 있었던 의미 있는 경험이었습니다.

참조문헌

김대기. (2025, December 14*). LG유플러스, '구글 AI 프로' 제휴 상품 출시…최대 반값 할인 혜택*. N뉴스. https://n.news.naver.com/mnews/article/009/0005605695

김수형. (2025, December 14). *"구글의 무서운 추격"…챗GPT 독주 끝내고 '2강 체제' 굳히나*. N뉴스. https://n.news.naver.com/mnews/article/055/0001316075

신동원. (2025, December 14). *유시민 "민주당, 뭐 하는지 모르겠다…지금 굉장히 위험" 경고.* N뉴스. https://n.news.naver.com/mnews/article/661/0000067125

이서현. (2025, December 14). *조국 "'계엄 사과' 25명 국힘 의원들, 뛰쳐나와 창당하라."* N뉴스. https://n.news.naver.com/mnews/article/666/0000090559

토픽 6급 어휘 통합 목록 TOPIK Level 6 Accumulate List. (n.d.). 한국어교육바. https://kleocean.com/%ED%86%A0%ED%94%BD-6%EA%B8%89-%EC%96%B4%ED%9C%98-%ED%86%B5%ED%95%A9-%EB%AA%A9%EB%A1%9D-topik-level-6-accumulate-list/

한자단어모음. (2013, December 22). 다움. https://wordbook.daum.net/open/wordbook.do?id=8975540

한희라. (2025, December 14). *해외 파생상품 투자하려면…모의거래 3시간 이수해야 가능.* N뉴스. https://n.news.naver.com/mnews/article/016/0002571717