[KDT] 기업맞춤형 AI+X 융복합 인재 양성 과정 - MBC Academy

뉴스 감성분석 기반 산업 리스크 조기 탐지 시스템

AI+X 융합 프로젝트 (뉴스 감성분석 & 주가 리스크 예측)

TEAM	PARTICIPANTS
2조	**김재호(팀장), 김민호, 박상용, 이정숙, 허남식**

Project Git 바로가기

목 차 (Contents)

01 프로젝트 개요

02 프로젝트 팀 구성 및 역할

03 프로젝트 수행 절차 및 방법

04 프로젝트 수행 경과 (기술 및 구현)

05 자체 평가 의견

01 프로젝트 개요

프로젝트 배경 및 목적

🎯 기획 의도 및 배경

- **리스크 관리의 중요성:** 경기 변동성 심화로 기업 및 산업 리스크의 조기 탐지 필요성 증대
- **정보의 비대칭성 해소:** 비정형 뉴스 데이터를 정량적 지표로 변환하여 객관적 모니터링 체계 구축
- **AI+X 융합 실습:** 파이썬 기반의 데이터 수집, 자연어 처리(NLP), 웹 서비스를 아우르는 풀스택 데이터 프로젝트 구현

💡 차별화 포인트

- **도메인 특화 모델:** 3대 타깃 산업(자동차, 건설, 헬스케어)에 최적화한 산업 분류 및 감성 분석 모델 적용
- **복합 리스크 지표:** 단순 감성 점수가 아닌, **거래량(Volume)**을 가중치로 활용하여 시장 파급력 측정
- **완전 자동화:** APScheduler를 활용한 수집-분석-적재-시각화의 Non-stop 파이프라인 구축

📌 기대 효과

뉴스 모니터링 업무의 효율화(정량화) 및 산업 리스크 발생 전조 증상 포착을 통한 선제적 경고 및 대응 지원

01 프로젝트 개요

개발 환경 및 워크플로

Language

Python 3.10+
SQL(Oracle Database)

Data Collection

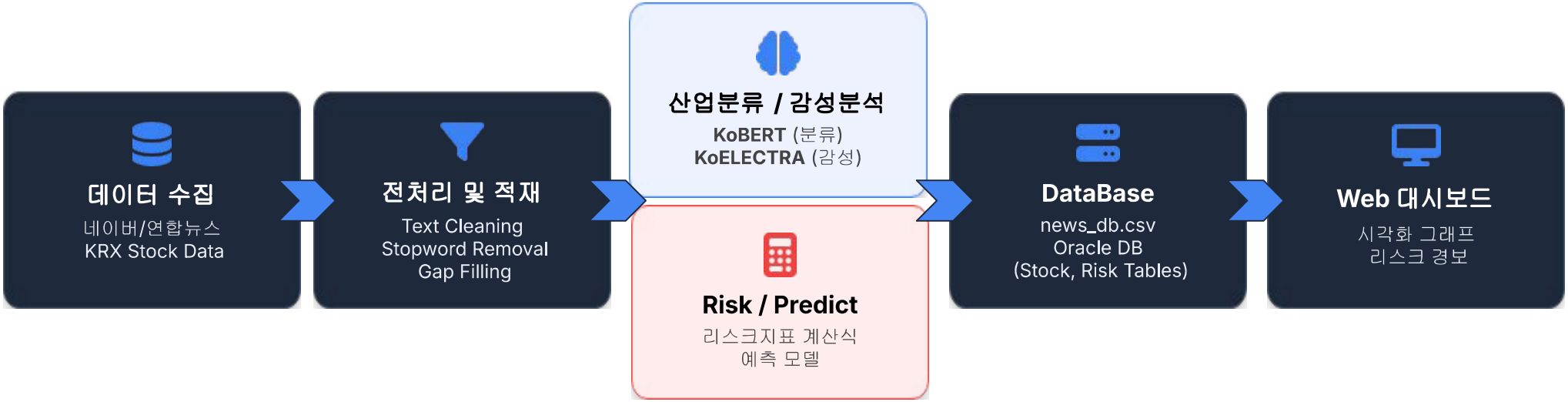
Selenium (Headless)
APScheduler

AI Model

KoBERT (산업 분류)
KoELECTRA (감성 분석)

Infra & Web

Oracle DB
Flask, Chart.js



02 프로젝트 팀 구성 및 역할

팀원별 역할 분담

이름	역할	담당 업무 상세	Contact
김재호	팀장(PM)	프로젝트 총괄, 시스템 아키텍처 설계, 연합뉴스 학습 데이터 KoBERT 모델링(산업 분류)	kimjh4462@gmail.com https://github.com/Jayk1220
김민호	팀원	주가 데이터(KRX) 수집, 리스크 지표(Risk Index) 산출 공식 수립	hora4444@naver.com https://github.com/hora4444
박상용	팀원	네이버뉴스 웹크롤링, APScheduler 자동화 로직 구현, 적재 파이프라인	thre3o2wo@gmail.com https://github.com/thre3o2wo
이정숙	팀원	Oracle DB 스키마 설계 및 구축, Flask 웹 서버 구축, DB 연동 API 개발, 데이터 시각화	jslee912@gmail.com https://github.com/jslee912
허남식	팀원	프로젝트 참고 논문 리서치, KoELECTRA 감성분석 모델 튜닝	huhnam59@gmail.com https://github.com/huhnam59

03 프로젝트 수행 절차 및 방법

수행 일정 및 주요 활동

	담당자	12/19	12/20	12/21	12/22	12/23	12/24	12/25	12/26	12/27	12/28	12/29	12/30	12/31	1/1	1/2	1/3	1/4	1/5
기획 및 설계																			
프로젝트 주제 선행 연구 조사	허남식																		
프로젝트 주제 확정 및 시스템 아키텍처 설계	김재호																		
데이터 수집																			
연합뉴스 크롤링	김재호																		
네이버 뉴스 크롤링	박상용, 이정숙																		
주가 데이터 수집 (KRX)	김민호																		
데이터 전처리																			
연합뉴스 기사 결측치 보정	김재호																		
네이버뉴스 기사 결측치 보정	박상용, 이정숙																		
주가 데이터 결측치 보정	김민호																		
뉴스 기사 산업별 분류 및 감성지수 산출	김재호																		
리스크 지표 생성	김민호																		
AI 모델링																			
산업 분류 모델 학습	김재호																		
감성 분석 모델 구축	허남식																		
주가 리스크 예측 모델링	김재호																		
시스템 구현																			
스케줄러 로직 구현	박상용																		
DB 스키마 구축 및 연동	이정숙																		
데이터 시각화 구현	이정숙																		
Flask 웹서비스 개발	이정숙																		
발표 준비																			
PPT 제작	박상용																		

04 프로젝트 수행 경과

① 데이터 수집 및 적재

학습 데이터 구축 (연합뉴스)

- **소스:** 연합뉴스 산업별 페이지 (명확한 카테고리 보유)
- **규모:** 약 2,800여 개 기사 (자동차, 건설, 헬스케어, 기타)
- **방식:** 본문이 없는 경우 제목으로 대체, 불용어 제거

👉 기사별 산업 분류 모델 학습

주기적 데이터 수집 (네이버뉴스)

- **소스:** 국내 11개 경제지 (25/10/01 ~ 현재)
- **규모:** 총 35만 건 이상 (일 평균 약 3,500건)
- **기술적 이슈:** API 사용 시 언론사 구분 적용 불가
  → **Selenium Headless Mode**로 해결
- **필터링:** 연예/스포츠 등 노이즈 기사 URL 기반 즉시 제거

일일 주가 정보 수집 (KRX INDEX)

- **소스:** 한국거래소 KRX Data Marketplace
- **데이터:** KRX 건설, KRX 자동차, KRX 헬스케어
  (25/09/30 ~ 현재) 일자별 종가, 거래량

실제 리스크 예측 시 입력 데이터 👈

04 프로젝트 수행 경과

② AI 모델 개요 및 성능

1. 산업 분류 모델 (KoBERT)

독립변수 : 기사 본문
타깃변수 : 산업분류(라벨인코딩)

📌 **KoELECTRA 모델도 유사한 성능을 보이나, 전체 성능 및 '해당없음' 분류에 KLUE-RoBERTa 근소하게 높음**

산업군	세부 지표	KLUE-RoBERTa (선정)	KoELECTRA	KcELECTRA
전체 성능	Total Accuracy	**0.9335**	0.9335	0.9124
	Macro F1	**0.8574**	0.857	0.8363
건설	F1-Score	0.6122	0.6122	0.6071
	Recall	0.5769	0.5769	0.6538
	Precision	0.6522	0.6522	0.5667
	Accuracy (이진)	**0.9667**	0.9667	0.9615
자동차	F1-Score	0.93	0.94	0.901
	Recall	0.949	0.9592	0.9286
	Precision	0.9118	0.9216	0.875
	Accuracy (이진)	**0.9755**	0.979	0.965
헬스케어	F1-Score	0.9333	0.9231	0.8989
	Recall	0.913	0.913	0.8696
	Precision	0.9545	0.9333	0.9302
	Accuracy (이진)	**0.9895**	0.9877	0.9842
[해당없음]	Accuracy (이진)	**0.9352**	0.9335	0.9142

04 프로젝트 수행 경과

② AI 모델 개요 및 성능

2. 감성 분석 모델 (KoELECTRA)

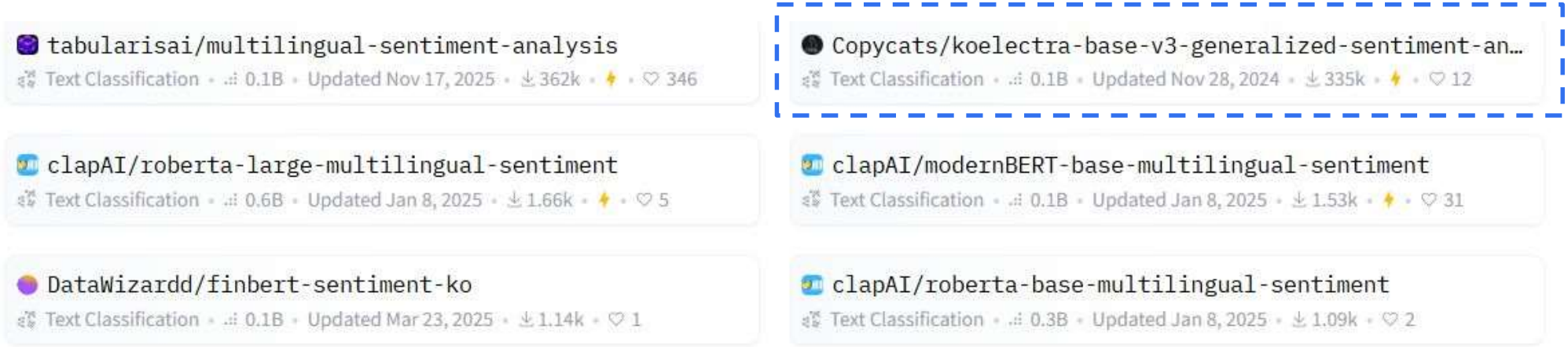
금융 도메인에 특화된 긍정/부정 점수 산출 (-1 ~ 1)

독립변수 : 기사 본문
타깃변수 : 감성점수 (정형데이터)

Base Model: (Huggingface)

koelectra-base-v3-generalized-sentiment-analysis

- 한국인이 제작한 모델 중 이용 수가 제일 높았음
- 단순 분류가 아닌 연속적인 수치화 가능



04 프로젝트 수행 경과

③ 리스크 지표 산출

📊 리스크 지표 (Risk Index) 산출 로직

$$Risk\_Index = Ave\_sentiment \times \ln(1 + \text{뉴스 본문 수} / \text{전체 뉴스 수}) \times \ln(1 + \text{거래량} / \text{전체 거래량})$$

🔑 변수별 의미 해석

항목	수식 내 변수	의미 및 역할
평균 감성	*Ave_sentiment*	뉴스 대화의 긍정/부정 수치(리스크의 성격, 시장의 방향성 결정)
뉴스 점유율	*뉴스 본문 수 / 전체 뉴스 수*	시장 이슈 중 해당 뉴스가 차지하는 관심의 집중도
거래 반응도	*거래량 / 전체 거래량*	실제 투자자들이 반응한 시장의 유동성(변동성) 비중
로그 정규화	$\ln(1+x)$	np.log1p를 통해 극단적인 값(outlier)의 영향을 보정

04 프로젝트 수행 경과

④ 주가 등락 예측 모델 실험

실험	모델	핵심 고도화 및 기술적 차별점	독립변수	종속변수	Accuracy	Precision	Recall	F1_score	MAE	R2	전략적 의의 (가설 및 성과)
1	DNN	기본 수치 데이터 기반 회귀 학습	감성지수, 종가, 거래량, 전일대비, 뉴스총량, 리스크지수, 기사/거래량 비중 (10개)	D+1 종가, D+2 종가 (연속형 수치)	-	-	-	-	174.8	-1.97	주가 수치 예측의 베이스라인 수립
2	DNN	One-Hot 인코딩 산업군 피처 추가	실험 1 변수 + 산업군 원-핫 인코딩 (건설, 헬스케어 등 카테고리 피처)	D+1 종가, D+2 종가 (연속형 수치)	-	-	-	-	160.24	-1.89	산업별 주가 변동 특성의 유효성 증명 (오차 감소)
3	DNN	Binary 방향성 분류 전환 (RMSprop)	감성지수, 종가, 거래량, 전일대비, 뉴스총량, 리스크지수, 기사/거래량 비중 (10개)	D+1 주가 상승 여부 (0 또는 1)	0.5625	0.5455	0.6316	0.5854	-	-	수치보다 '방향(상승/하락)' 예측이 현실적
4	DNN	Risk×Sentiment 상호작용 & 노이즈 주입	실험 3 변수 + 리스크×감성 상호작용, 리스크×거래량 상호작용 (12개)	D+1 주가 상승 여부 (0 또는 1)	0.625	0.6	0.75	0.6667	-	-	**Risk Index가 주가의 강력한 선행지표임을 입증**
5	DNN	리스크 변화량 및 임계점(0.7) 피처 확장	실험 4 변수 + 리스크 변화량 (Δ), 임계점(0.7) 돌파 여부 (14개)	D+1 주가 상승 여부 (0 또는 1)	0.5	0.5	0.4375	0.4667	-	-	과적합(Overfitting) 확인
6	**Random Forest**	Ensemble 머신러닝 교차 검증	실험 4와 동일 (리스크 상호작용 피처 포함 12개)	D+1 주가 상승 여부 (0 또는 1)	0.625	0.6154	0.5714	0.5926	-	-	**비선형 관계 포착을 통해 예측 정밀도(신뢰도) 극대화**
7	DNN	Batch Normalization & Adam 최적화	실험 4와 동일 (변수는 동일하나 배치 정규화 층을 통해 데이터 분포 조정)	D+1 주가 상승 여부 (0 또는 1)	0.5938	0.5789	0.6875	0.6286	-	-	학습 안정화 기법을 통한 모델의 범용적 성능 확보

04 프로젝트 수행 경과

④ 주가 등락 예측 모델

주가 등락 여부 예측 목적 학습 모델 (RandomForest)

Accuracy : 0.625

Precision : 0.6154

Recall : 0.5714

F1_score : 0.5926

📌 **Random Forest를 사용한 이유:**

- **노이즈와 과적합(Overfitting) 방지**
- **accuracy가 동일함에도, Precision이 RF가 더 높음.**
- **주식에서는 허위 상승 신호에 속아 원금을 잃는 리스크를 최소화하는 것이 가장 중요하기에 Precision의 수치가 더 중요하다고 판단**

⑤ DB 구성

* 글씨색 컬럼 : 웹크롤링 데이터
* 배경색 컬럼 : 대시보드 출력 데이터

1. NEWS (csv)

컬럼명	설명
NDATE	기사 일자 및 시간
TITLE	기사 제목
CONTENT	기사 본문
OID	언론사 구분
LINK	기사 원문 링크
INDUSTRY	**산업분류 결과**
SENT_SCORE	**감성분석 점수**

2. STOCK (Oracle DB)

컬럼명	설명
SDATE	기준일자
MARKET_INDEX	KRX 주가지표 구분
CLOSE	기준일 종가
CHANGE	전일대비 종가 차이
VOLUME	거래량

3. RISK (Oracle DB)

컬럼명	설명
RDATE	기준일자
INDUSTRY	산업 분류
AVE_SENTIMENT	**기준일/산업 감성분석 점수 평균**
TOTAL_NEWS	**뉴스기사 총량**
ARTICLE_RATIO	**산업별 기사 비율**
TOTAL_VOLUME	**전체 거래량**
TRADE_VOLUME_RATIO	**전체 대비 산업별 거래량 비율**
RISK	**리스크지표**
PREDICT	**주가등락 예측**

⑥ 서비스 구현

자동화 (APScheduler)

- 10분 단위: 뉴스 크롤링 및 AI 분석 (Gap Filling)
- 24:00: KRX 주가 마감 데이터 수집
- 00:30: 일일 리스크 지표, 등락 예상 최종 산출 및 DB 적재

웹 구현 (Flask)

- Oracle DB의 STOCK, RISK 테이블 연동
- Chart.js를 활용한 주가 vs 리스크 비교 그래프 시각화
- 산업별 등락 예상(Up/Down) 아이콘 출력

⑦ 대시보드 시연 (Demo)

05 자체 평가 의견

프로젝트 성과 및 제언

✅ 성과 및 특장점

- **정량적 모니터링:** 담당자의 직관에 의존하던 뉴스 분석을 통계적 지표 기반으로 전환
- **데이터 파이프라인 완성:** 수집부터 시각화까지 전 과정을 끊김 없이 자동화하여 실무 적용 가능성 확인
- **높은 분류 정확도:** 세 가지 타깃 산업 분야 모두에서 96% 이상의 분류 성능 확보

⚠ 아쉬운 점 & 한계

- 단순 등락(Up/Down) 외에 구체적인 주가 수치 예측은 어려움
- 기간 내 구현을 위해 산업군을 3개로 한정했던 점
- 약 3개월의 학습 내용으로는 예측 정확도를 높이는 데 한계

🚀 향후 개선 방향

- LLM (ChatGPT 등) 도입을 통한 리스크 발생 원인 요약 기능 추가
- 각 산업 분야 특화 말뭉치 추가 학습으로 분류 성능 개선
- 산업군 추가 분류로 타깃 산업군 범위를 넓혀 리스크지표 수식상 각 산업군 기사량이 타 산업군의 리스크 지표에 미치는 영향 완충
- 더 폭넓은 크롤링을 통한 과거 데이터 추가 보충으로 주가등락 예측모델 학습 결과 개선

End of Presentation

참고문헌

1. 강두원, 유소엽, 이하영 and 정옥란. (2022). 뉴스 감성 분석을 이용한 딥러닝 기반 주가 예측에 대한 연구. 한국컴퓨터정보학회논문지, 27(8), 31-39. [링크]