뉴스데이터 전처리 결과 보고서

1. 목적및배경

최신 뉴스 데이터를 수집하고 감정 분석을 수행하여 주가 예측 및 투자 심리 지표 개발을 목표로 함.

CurrentsAPI 를 사용하여 뉴스 데이터 수집

감정 분석 모델(BERT, LSTM, VADER, XGBoost) 학습 및 비교
DeepSeek-r1 모델을 활용한 감정 분류 결과 사유 추론
Exaone 3.5 모델을 이용한 한국어 번역

2. 데이터 개요

2.1 데이터 수집일자

최신 뉴스 데이터 기준으로 실시간 수집 및 분석 수행

.

2.2 데이터 양

총 4,846 개의 뉴스 데이터

감정 분석 결과 포함

2.3 대상 소스

API: CurrentsAPI
AI 모델: BERT, LSTM, VADER, XGBoost, DeepSeek-r1(Ollama), Exaone 3.5
언어 및 라이브러리: Python, requests, transformers, ollama, exaone-sdk

3. 데이터 전처리 목표

수집된 뉴스 데이터를 감정 분석 및 번역하기 위해 아래와 같은 전처리 과정 수행.

3.1 데이터 필터링

주제(business) 관련 뉴스만 필터링

중복 기사 및 불필요한 광고/스팸 기사 제거
감정 분석에 필요 없는 단어(ex:’의’, ‘그리고’, ‘그’) 제거

3.2 데이터 정제

특수문자 및 HTML 태그 제거

뉴스 제목 및 본문만 추출

불필요한 공백 정리

4. 데이터 처리 단계

4.1 감정 분석 모델 학습 및 평가

사용 모델: BERT, LSTM, VADER, XGBoost

모델 평가 기준: 정확도 및 혼동 행렬(Confusion Matrix) 비교
결과: BERT 모델이 가장 높은 정확도 기록

4.2 감정 분석 결과 사유 추론

Ollama DeepSeek-r1 모델을 사용하여 감정 분류의 사유 추론

4.3 뉴스 번역 및 데이터 변환

Exaone 3.5 모델을 사용하여 한국어 번역 수행

현재는 Exaone3.5:7.8B 모델을 사용하여 결과 추론 및 번역

4.4 데이터 저장

CurrentsAPI 에서 수집한 JSON 데이터를 정제 후 저장

5. 향후 사용 계획

주가 예측 모델에서 감정 분석 데이터로 활용

감정 분석 기반 투자 심리 지표 개발