**비정형 데이터 활용 분석 프로젝트 기획안**

**2024년 5월 22일**

**과정명: 금융&마케팅 데이터 분석(Python)**

| 프로젝트 주제 | 머신러닝 기술과 소셜데이터를 활용한 주가 어드바이저 개발 |
| --- | --- |
| 팀 명 | SocialStock |
| 팀 원 | 노재민(팀장), 박찬우, 김은지, 유지원 |
| 프로젝트 내용  해결하고자 하는 문제  최종 산출물의 청사진 | 본 프로젝트는 미국 반도체 업종에 집중 투자하는 ETF인 ‘SOXL’의 주가데이터와 소셜데이터를 심층 분석하여 투자할 때 유용한 지표를 제공하는 것을 목표로 함. 최종적으로 변동성이 큰 반도체 시장에서 투자 결정을 돕기 위해 특정 시점을 입력시 ‘SOXL’에 대하여 5단계(강력매도, 매도, 중립, 매수, 강력매수)로 나눠진 의견을제시하는 걸 목적으로함. |
| 프로젝트 추진 일정 | * 5/20(월) – 브레인스토밍, 아이디어 선정, 기획안 및 개요,키워드선정, 데이터수집 * 5/21(화) 데이터 전처리 , 모델링시작(60%이상 성능달성시 종료) * 5/27(월) – 데이터 전처리 및 모델링 완료 * 5/28(화) ~ 5/30(목) – 데이터 시각화 (스트림릿) * 5/30(목) – 마무리 작업, PPT 작성 완료 * 5/31(금) – 발표 연습 및 최종점검, 경진대회 |
| 프로젝트 수행 방향  수행 방법/도구 | * **키워드 선정** * 검색 키워드 추출: 소셜 데이터에서 주가와 상관관계가 높은 검색키워드 선정 * 수행 방법/도구: 나스닥 공식 홈페이지 에서 제공하는 SOXL 구성성분을 활용하여 종목내의 구성성분을 찾고 해당 티커명을 엑셀파일로 저장 * **데이터 수집(deadline 5/21**) * 소셜 데이터 수집: Investing.com, reddit, 네이버 종목토론실에서 ‘SOXL’, ‘NVDA’ 등 관련된 키워드를 포함한 데이터 수집 * 주가 데이터 수집:’SOXL’의 과거 주가 데이터를 수집 * 수행 방법/도구: Selenium, BeautifulSoup4,listly * **데이터 분석 및 전처리(deadline 5/27)** * 상관관계 분석: 소셜 데이터에서 추출된 키워드와 감정 점수, 주가 변동 간의 상관관계를 분석 * 데이터 전처리: 수집된 데이터를 정제하고, 필요에 따라 토큰화, 불용어 제거 등의 NLP기법을 사용하여 전처리 * 수행 방법/도구: pandas, numpy, sklearn, transformer,matplotlib, seaborn * **빅데이터 모델링 및 예측(deadline 5/27)** * 예측 모델 개발: 소셜 데이터와 주가 데이터를 결합하여 주가 모델 예측 * 모델 평가: 예측 모델의 성능을 평가, 필요에 따라 모델 튜닝 * 수행 방법/도구: pandas, numpy, transformer, Logistic Regression, Random Forest Classifier, KNN, Support Vector Classifier, Gradient Boosting, Decision Tree Classifer, Bagging, RNN, LSTM, XGBClassifier Classifier, lgbm * **결과 시각화(deadline 5/30)** * 결과 시각화: 분석 결과를 보다 직관적으로 이해할 수 있도록, Streamlit을 활용하여 시각화 * 주식의 주가 변동 추이, 소셜 미디어에서 언급된 키워드와 주가의 상관관계, 그리고 매수 또는 매도를 추천하는 의견 전달 * 수행 방법/도구: streamlit , matplotlib, seaborn * **발표 준비(deadline 5/31 오전)** * 피피티 제작및 발표 대본 정리 |