Final Project Report(영어분석을위한기계학습)

경제학 전공 201903508 채승우

프로젝트명: Reddit Stock Sentiment Analyzer

1. 서론: 문제 정의 및 프로젝트 목표

1.1. 문제제기

미국 주식 시장은 투자자의 심리에 수익률이 달라진다. Reddit의 주식 갤러리와 같은 커뮤니티에는 하루에도 수천, 수만 개의 게시글이 올라오는데 투자자가 이 방대한 양의 텍스트를 일일이 읽고 그 기업에 대한 전반적인 심리를 파악하는 것은 물리적으로 불가능하다.

1.2. 프로젝트 목표

본 프로젝트는 기업에 대한 다른 투자자의 심리점수를 구축하는 것을 목표로 합니다. Reddit의 실시간 게시글을 수집하고, 머신러닝 모델을 통해 해당 글이 Bullish인지 Bearish인지를 자동으로 분류하여, 투자자가 직관적으로 특정 기업에 대한 투자자의 심리를 파악할 수 있도록 돕는 모델을 개발하고자 한다.

2. 진행 과정

본 프로젝트는 머신러닝 파이프라인의 정석적인 절차를 따라 다음과 같이 진행되었다.

2.1. 데이터 수집

- **방법:** Reddit API의 접근 제한 문제를 인지하여 다른 방법을 시도하였고 실시간성을 확보의 장점이 있는 **Reddit RSS Feed**를 활용하였다.

- **대상:** 주요 주식 서브레딧 (r/stocks, r/wallstreetbets, r/investing)

- **결과:** 게시글의 제목(Title)과 본문(Selftext) 데이터를 실시간으로 수집하는 파이프라인을 구축하였다.

2.2. 데이터 전처리 및 레이블링

- **레이블링:** 수집된 데이터에 대한 지도 학습을 규칙 기반 감성 분석 라이브러리인 VADER를 사용하여 진행하였고 레이블(Bullish/Bearish)을 자동으로 부여한다.

- **전처리:** 영문 텍스트 클리닝 후, 기계 학습을 위해 TF-IDF Vectorizer를 사용하였고 상위 1,000개의 단어 특징을 추출하였다.

2.3. 모델링

- **모델 선택:** 고차원 희소 데이터 처리에 효율적이고 학습 속도가 빠른 SGD Classifier (Logistic Regression)를 사용하였다.

- **학습 설정:** Log Loss 함수를 사용하여 확률적 경사 하강법으로 학습을 진행하였고, Train/Validation/Test 셋을 8:1:1로 엄격히 분할하여 과적합 문제를 해결하였다.

3. 모델 학습 결과 및 분석

3.1. 학습 성능

모델 학습 과정에서 Loss(손실)는 꾸준히 감소하고 Accuracy(정확도)는 상승하는 안정적인 수렴 곡선을 보였다.

3.2. 최종 평가 및 한계점 분석

최종 Test Set(25개 샘플)을 돌린 결과, Accuracy 100%, F1-Score 100%를 달성했다.

한계 및 배운점**:** 100%라는 성능 수치가 의아하게 여겨져서 분석하였는데 2가지의 가능성이 제기되었다. (1) 모델을 학습시킬 때 사용하였던 데이터와 Vader에 사용하였던 데이터가 동일하였기 때문, (2) 테스트 데이터셋의 크기(Sample Size)가 작았기 때문으로 분석된다. 이는 단순히 코드를 잘못 짰기 때문으로 평가된다.

또한 알아서 데이터를 크롤링하여 현재 특정 기업에 대한 심리 점수를 매기고 싶었지만 데이터를 자동으로 크롤링하게 하여 전부를 분석하게 하는 코드를 짜지 못하여 아쉽게 생각한다.

4. 실제 사용 예시 (Real Usage)

4.1. 사용 시나리오

사용자가 Reddit 본 문장을 입력하면, 모델이 즉시 이를 해석하고 분석하여 Bullish/Bearish 여부를 수치와 함께 출력한다.

4.2. 추론 결과

다음은 실제 Reddit 게시물을 입력했을 때의 모델 반응이다.

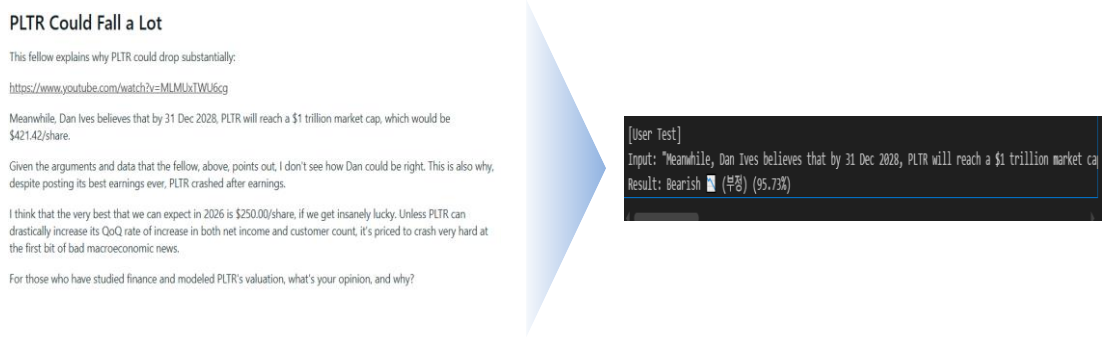
Case 1: 긍정적 뉴스 (호재)

- **입력:** *"Nvidia earnings smashed expectations! AI demand is unstoppable."*

- **결과:** 🚀 **Bullish** (예측 성공)

Case 2: 부정적 뉴스 (악재)

- **입력:** *"Inflation data is worse than expected, markets represent a huge risk."*
- **결과:** 📉 **Bearish** (예측 성공)



5. 결론 및 향후 계획

5.1. 프로젝트의 의의

이번 학기 프로젝트를 통해 막연하게만 느껴졌던 비정형 텍스트 데이터의 수집부터 머신러닝 모델 배포까지의 전 과정을 직접 경험해 볼 수 있어 뜻깊은 시간이었다. 특히, 단순히 모델을 돌리는 것을 넘어 주의할 점을 공부할 수 있어 좋았다.

5.2. 향후 발전 계획

현재 버전은 사용자가 텍스트를 수동으로 입력해야 한다는 한계가 있습니다. 이를 발전시켜 다음과 같은 기능을 추가할 계획이다.

- **실시간 크롤링 통합:** 사용자가 기업 티커(예: TSLA)만 입력하면, 자동으로 최신 글 100개를 긁어와서 평균 감성 점수를 보여주는 기능
- **데이터 규모 확대:** Kaggle 등의 대규모 데이터셋을 추가 학습하여 모델의 일반화 성능 강화
- **딥러닝 모델 도입:** BERT나 FinBERT와 같은 트랜스포머 기반 모델을 도입하여 문맥을 더 깊이 있게 이해하는 모델 개발

본 프로젝트는 저만의 'AI 투자 비서'를 만들기 위한 첫걸음이었으며, 앞으로도 계속 발전시키도록 하겠습니다.