Project Report: Sentiment Analysis and Stock Price Prediction

Overview

The objective of this project is to explore the connection between public sentiment and stock prices. By utilizing sentiment analysis on news headlines and financial data, we aim to build a predictive model that can inform trading decisions. The project is divided into several key stages, including data collection, sentiment analysis, stock data retrieval, sentiment-adjusted moving averages, and signal generation for evaluating trading strategies.

Project Flow

1. **Data Collection:**
    - **Web Scraping of News Headlines:** Using Selenium, we scraped news headlines related to specific stocks from Yahoo Finance. This involved dynamically loading pages and extracting relevant data within a 12-month timeframe.
    - **API Fetching from New York Times:** To enhance our dataset, we fetched additional news articles from the New York Times using their archive API, focusing on articles related to Apple Inc.
2. **Data Preprocessing:**
    - **Date Conversion:** We converted scraped news dates, often in relative formats like "2 hours ago" or "yesterday," to actual dates.
    - **Merging Datasets:** We merged the Yahoo Finance and New York Times datasets to ensure a comprehensive collection of news headlines for our target stocks.
3. **Sentiment Analysis:**
    - **VADER Sentiment Scoring:** Each headline was analyzed using the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool to assign a polarity score, categorizing the sentiment as positive, negative, or neutral.
4. **Stock Data Retrieval:**
    - **Yahoo Finance API:** We retrieved stock data for selected companies using the yFinance API, spanning from July 1, 2018, to June 14, 2024. This included open, high, low, close, adjusted close prices, and volume.
5. **Data Integration:**
    - **Merging Stock and Sentiment Data:** We integrated the sentiment scores with stock data based on the date and stock ticker.
6. **Feature Engineering:**
    - **Sentiment-Adjusted Moving Averages:** We developed a sentiment-adjusted moving average (SMA) to incorporate sentiment scores along with traditional price data.
    - **Preprocessing Headlines:** Headlines were preprocessed using tokenization, lemmatization, and TF-IDF vectorization for model training.
7. **Model Training and Evaluation:**
    - **Random Forest Classifier:** A Random Forest classifier was trained to predict sentiment labels. The model was evaluated using cross-validation and test sets.
    - **Signal Generation:** Buy/sell signals were generated for trading decisions based on sentiment scores and moving averages.
    - **Performance Metrics:** We evaluated the trading strategy using metrics such as the Sharpe Ratio, win ratio, and overall portfolio value.

8. **Visualization:**
    - **Plotting Results:** We visualized stock price, sentiment scores, buy/sell signals, and portfolio value over time to assess the strategy's effectiveness.

Detailed Project Flow

Data Collection

- **Web Scraping with Selenium:** We used Selenium WebDriver to scrape news headlines from Yahoo Finance by simulating user interactions, collecting news related to specific stock tickers over a 12-month period.
- **Fetching Data from New York Times API:** Using the New York Times archive API, we retrieved additional articles mentioning Apple Inc., filtered by relevant keywords.

Sentiment Analysis

- **Using VADER for Sentiment Analysis:** Each headline was analyzed with VADER to determine its sentiment, providing a polarity score that classified the sentiment as positive, negative, or neutral.

Stock Data Retrieval and Integration

- **Fetching Stock Data:** Stock price data for selected companies was retrieved using the yFinance API, including metrics like open, high, low, close, adjusted close prices, and trading volume.
- **Integrating Sentiment and Stock Data:** Sentiment scores were merged with stock data based on date and stock ticker, allowing us to analyze sentiment's impact on stock prices.

Feature Engineering and Model Training

- **Sentiment-Adjusted Moving Averages:** We created sentiment-adjusted moving averages (SMAs) by incorporating sentiment scores into traditional moving average calculations to better capture sentiment's influence on stock prices.
- **Random Forest Classifier for Sentiment Prediction:** A Random Forest classifier was trained to predict news headline sentiment, evaluated using cross-validation techniques.

Signal Generation and Performance Evaluation

- **Trading Signals Based on Sentiment-Adjusted Moving Averages:** Trading signals were generated based on sentiment-adjusted moving averages, indicating potential buy or sell actions to optimize trading performance.
- **Performance Metrics:** We evaluated the strategy using metrics like the Sharpe Ratio, win ratio, and overall portfolio value, assessing the sentiment-adjusted strategy's effectiveness.

Conclusion

This project demonstrates the potential of integrating sentiment analysis with stock price prediction. By analyzing news headline sentiment and incorporating it into traditional moving averages, we developed a sentiment-adjusted trading strategy. While the strategy's performance evaluation will reveal its effectiveness, the approach offers a promising direction for enhancing trading algorithms with sentiment analysis.

Future Work

- **Model Optimization:** Further tuning and optimization of the Random Forest model and other machine learning algorithms.
- **Real-time Sentiment Analysis:** Implementing a real-time system to analyze live news feeds and adjust trading strategies accordingly.
- **Exploration of Other Sentiment Analysis Tools:** Experimenting with advanced NLP models like BERT for more accurate sentiment predictions.
- **Expansion to Other Stocks and Markets:** Extending the analysis to a broader range of stocks and exploring other financial markets.

By combining sentiment analysis with financial data, this project lays the groundwork for developing more sophisticated and responsive trading strategies that leverage natural language processing and machine learning.