Predicting Stock Movements Using Sentiment Analysis and Machine Learning

By RIDA SHAHWAR

1. Introduction

This project focuses on analyzing the stock price movements and predicting future trends for five stocks: ONCO, CNEY, TNXP, APLD, and KTTA. The analysis combines historical stock data, sentiment analysis from news headlines, time-series decomposition, and machine learning models to forecast stock prices. Data was gathered from early 2021 to October 2024.

2. Data Extraction and Preparation

Libraries Used:

- yfinance for extracting historical stock data.
- pandas for data manipulation and handling.
- requests and BeautifulSoup for web scraping to gather news data related to each stock.
- matplotlib and plotly for data visualization.

Stock Data:

Historical stock data for the tickers ONCO, CNEY, TNXP, APLD, and KTTA were extracted using yfinance covering the period from early 2021 to October 2024.

News Data:

For sentiment analysis, news headlines were fetched using yfinance and BeautifulSoup, which extracted relevant news articles for each stock. The sentiment analysis was performed on these headlines to gauge the public perception and its potential impact on stock prices.

3. Sentiment Analysis

Purpose:

The objective of sentiment analysis was to classify news articles related to the five stocks as positive, neutral, or negative. This was done using sentiment analysis libraries TextBlob and VADER (Valence Aware Dictionary for Sentiment Reasoning).

Methodology:

- TextBlob was used to calculate the polarity score of each news headline. The scores ranged from -1 (negative sentiment) to +1 (positive sentiment).
- VADER analyzed each headline to provide a compound sentiment score and categorize the sentiment as positive, negative, or neutral.

Output:

A DataFrame `df_news_sentiment` was created, containing:

- Stock Ticker
- Headline
- Polarity (TextBlob)
- Compound Score (VADER)
- Sentiment Classification (Positive, Neutral, Negative)

Results:

The average sentiment for each stock:

- ONCO: 0.25
- CNEY: 0.34
- TNXP: 0.0011
- APLD: 0.03
- KTTA: 0.27

4. Time Series Analysis

Decomposition:

Time series decomposition was performed on the closing prices of each stock. Using the seasonal_decompose function, the series was broken down into three components: trend, seasonality, and residuals**.**

Stationarity Check:

The Augmented Dickey-Fuller (ADF) test was applied to each stock to assess stationarity. The results showed that ONCO and TNXP were stationary (p-value < 0.05), whereas CNEY, APLD, and KTTA were not**.**

Autocorrelation Plots:

Autocorrelation (ACF) and partial autocorrelation (PACF) plots were used to identify patterns in the time series for ARIMA modeling.

5. Machine Learning Predictions

Model Used:

A Random Forest Classifier was employed to predict whether the stock price would increase the following day (binary classification: 1 for increase, 0 for no increase). The model used both stock price data and sentiment features.

Features:

- Moving Averages (10-day and 20-day)
- Sentiment scores
- Previous day's closing price.

Model Evaluation:

The Root Mean Squared Error (RMSE) was calculated for each stock:

- ONCO: 17.44
- CNEY: 5.32
- TNXP: 2,197,765.87 (extreme outlier)
- APLD: 0.36
- KTTA: 5.32.

Accuracy:

Model accuracy varied across stocks, with Random Forest achieving different levels of success due to varying volatility and noise in the stock prices.

6. Visualization

Sentiment Distribution:

A bar chart was created to visualize the sentiment distribution by stock. Each stock had different proportions of positive, neutral, and negative headlines.

Time Series Plots:

Time series plots of closing prices over time were generated for each stock to highlight trends and fluctuations in the market(vertopal.com_Stock_Anal…).

Prediction Accuracy:

Scatter plots comparing actual versus predicted closing prices for each stock were created to assess model performance(vertopal.com_Stock_Anal…).

7. Conclusion

This project offers a detailed analysis of stock performance, combining sentiment analysis, statistical methods, and machine learning models. The findings highlight the importance of sentiment in driving stock prices, especially in volatile stocks like TNXP. While machine learning models showed varying prediction accuracy, integrating sentiment data improved predictions.

Future work could explore refining the feature set, trying different models like LSTM (Long Short-Term Memory), and improving sentiment analysis with more granular data.

NOTE: There are two python files Stock_Analysis1 and Stock _Analysis2