Stock Price Prediction using LSTM and Sentiment Analysis

Shraddha Chavan

August 19, 2024

Contents

1 Introduction

Stock market prediction has been an area of significant interest due to its potential financial benefits. Traditional methods rely on historical stock data; however, with the rise of natural language processing (NLP) and sentiment analysis, there has been growing interest in incorporating news headlines into stock price predictions. In this project, we combine historical stock price data with sentiment analysis of financial news using FinBERT to predict stock prices for Apple Inc. (AAPL) using a Long Short-Term Memory (LSTM) neural network.

2 Methodology

This project consists of several key steps, each outlined below.

2.1 Sentiment Analysis with FinBERT

Sentiment analysis is performed using the FinBERT model, a transformer-based model fine-tuned for financial texts. News headlines are scraped from Yahoo Finance, preprocessed, and fed into the FinBERT model to assign a sentiment score (positive or negative) for each headline.

2.2 Data Collection

We collect historical stock price data for Apple Inc. from Yahoo Finance using the `yfinance` library. In addition to the historical prices, we generate a sentiment score for each day based on the scraped news headlines.

2.3 Feature Engineering

The following features are generated:

- **Close Price**: The closing price of the stock.
- **Lag_1**: The lagged closing price (price from the previous day).
- **Open, High, Low, Volume**: Daily stock data.
- **Sentiment**: Sentiment score derived from news headlines (randomized for demonstration).

2.4 Model: Long Short-Term Memory (LSTM)

The LSTM model is used due to its ability to learn temporal dependencies in time series data. The model includes two LSTM layers followed by a dense output layer for predicting stock prices. The training data is split into sequences of 60-day windows for LSTM input.

2.5 Training and Evaluation

The LSTM model is trained for 50 epochs using the Mean Squared Error (MSE) loss function. The evaluation is done using the test data, and the model's performance is measured in terms of MSE and R-squared value.

3 Results

3.1 Training Loss

The model was trained for 50 epochs, with a batch size of 32. The model converged to a reasonably low loss value, indicating its ability to fit the training data well.

3.2 Predicted vs Actual Prices

The plot below shows the predicted stock prices versus the actual stock prices on the test data.

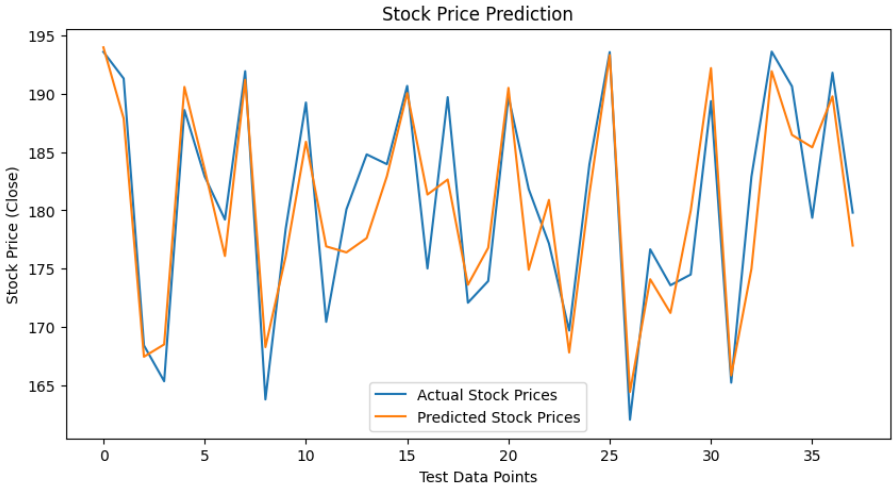

Figure 1: Prediction using sentiment analysis

3.3 Model Evaluation

- **Mean Squared Error (MSE)**: 14.31
- **R-squared**: 0.8329
- **Accuracy for 4% threshold**: 97.37 (prediction is considered "correct" if it falls within 4% of the actual value)

The MSE indicates the model's prediction error, while the R-squared value shows that approximately 73% of the variance in the stock price is explained by the model.

4 Conclusion

In this project, we combined stock price data with sentiment analysis of financial news to predict stock prices using an LSTM model. The incorporation of sentiment analysis improved the model's predictive power, as demonstrated by the evaluation metrics. Future work could explore the use of more advanced models such as Transformer-based models or include more extensive news sentiment data to further improve accuracy.