

1. Choice of Dataset

Alpha Vantage API:

We have chosen to utilize the Alpha Vantage API for real-time and historical data on stocks. The comprehensive dataset includes various metrics like open, high, low, close prices, and volume, enabling us to analyze stock price movements accurately. We aim to predict the future prices of selected stocks, and this API provides us with reliable and up-to-date data for our machine learning model.

Twitter Data:

In addition to the stock data, we will incorporate sentiment analysis from Twitter posts relating to the specific stocks we are analyzing. We believe that public sentiment on social media can significantly influence stock prices. By scraping Twitter for real-time tweets and utilizing sentiment analysis, we aim to gauge the public sentiment and factor it into our stock price prediction model.

2. Methodology

a. Data Preprocessing:

- Alpha Vantage Data: We will clean and preprocess this data, handling any missing values, normalizing the data, and possibly engineering features to make the dataset conducive for training our machine learning model. They provide a phenomenal amount of data on each stock and since many years are available, it'll be easy to create our sets as we can fetch the data and also the price it actually ended up trading at.
- Twitter: Tweets will be cleaned to remove any irrelevant information, and sentiment analysis will be performed to classify the sentiment of each tweet as positive, negative, or neutral.

b. Machine Learning Model:

We plan to implement a hybrid model integrating time series forecasting and sentiment analysis. The primary model for stock price prediction could be a Long Short-Term Memory (LSTM) network, given its efficacy in sequence prediction problems like stock prices.

- Pros: Capable of capturing long-term dependencies, handling sequences, and modeling non-linear relationships.
- Cons: Can be computationally intensive and requires optimal tuning to prevent overfitting or underfitting.

We are also thinking about exploring other models like ARIMA and comparing their performance.

c. Evaluation Metric:

For our regression problem, we'll primarily use metrics like Mean Squared Error (MSE) to evaluate the performance of our model. We will also consider additional metrics like Mean Absolute Error (MAE) for a more comprehensive evaluation.

3. Application

Our model will be integrated into a user-friendly web application. Users can input the stock ticker symbol, and optionally, a date range for historical data analysis.

- User input: Users provide the stock ticker symbol through a search bar, and the application fetches real-time data from the Alpha Vantage API and Twitter.
- Output: The web app will display the predicted stock price, supplemented with sentiment analysis results.

The prediction and sentiment analysis will equip users with insightful data to make informed decisions on their stock investments. We aim for a responsive design to ensure accessibility and convenience for all users.