Stock Prediction AI



Kicki Nocoj Bici

EC Utbildning

AI projektarbete

20241031

# Abstract

Predicting future stock prices has been a long-standing challenge in financial markets, requiring sophisticated methods to analyze complex and ever-changing datasets. This project focuses on building an Artificial Intelligence (AI) model aimed at predicting stock prices by leveraging a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) approach. The model pipeline integrates diverse data sources, including technical indicators, macroeconomic data, and sentiment analysis derived from company news. Due to the large volume of data, the project focused specifically on collecting stock data and building a model based on Nasdaq tech stocks, including 'AAPL', 'MSFT', 'GOOGL', 'TSLA', 'NVDA', 'META', 'AMZN', 'NFLX'. The use of modern machine learning techniques, such as hyperparameter optimization and backtesting, ensures that the model is well-suited to capture market trends and provide reliable predictions. This report presents the methodology used for data collection, feature engineering, model development, and evaluation, culminating in a comprehensive backtesting procedure to validate the model's performance.

Innehållsförteckning

[Abstract 2](#_Toc181104030)

[1 Introduction 1](#_Toc181104031)

[2 Theory 2](#_Toc181104032)

[2.1.1 Time-series models 2](#_Toc181104033)

[2.1.2 Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) 2](#_Toc181104034)

[2.1.3 Sentiment analys 2](#_Toc181104035)

[2.1.4 The hybrid 2](#_Toc181104036)

[3 Metod 4](#_Toc181104037)

[3.1 Data Collection 4](#_Toc181104038)

[3.2 Feature Engineering 4](#_Toc181104039)

[3.3 Sentiment Analysis 4](#_Toc181104040)

[3.4 Model Training 5](#_Toc181104041)

[3.5 Evaluation 5](#_Toc181104042)

[3.6 Backtesting 5](#_Toc181104043)

[4 Resultat och Diskussion 6](#_Toc181104044)

[5 Conclusion 7](#_Toc181104045)

[6 Reflections 8](#_Toc181104046)

[Appendix A 9](#_Toc181104047)

[References 10](#_Toc181104048)

# Introduction

The stock market is characterized by high volatility and unpredictability, influenced by numerous factors such as economic conditions, company performance, and investor sentiment. Accurately predicting stock prices has significant implications for investors, financial analysts, and institutions looking to make informed decisions. Traditional methods for stock price prediction often rely on historical data and statistical techniques. Still, with advancements in Artificial Intelligence (AI) and machine learning, more sophisticated models are being developed to improve prediction accuracy.

This project aims to build an AI model that predicts future stock prices by integrating data from multiple sources, such as stock prices, macroeconomic indicators, and sentiment from news articles. Due to the large volume of data and practical constraints, the project focused on Nasdaq tech stocks, specifically 'AAPL', 'MSFT', 'GOOGL', 'TSLA', 'NVDA', 'META', 'AMZN', 'NFLX'. By employing a hybrid CNN-LSTM model, the project seeks to exploit both spatial patterns in time-series data and temporal dependencies, offering a more nuanced understanding of market behavior. The combination of Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) networks for sequence learning allows the model to identify patterns and trends that traditional statistical methods might miss.

The overall pipeline for the project consists of several critical stages, including data collection, feature engineering, sentiment analysis, model training, and backtesting. Each of these steps is designed to enhance the predictive capabilities of the model while ensuring robustness and reliability in real-world scenarios. This report provides a detailed overview of the entire process, from initial data collection to final model evaluation, highlighting the methods and tools used at each stage to achieve the best possible results.

# Theory

The theory behind predicting stock prices using machine learning involves understanding both the underlying financial concepts and the technical implementation of various algorithms. Stock prices are influenced by numerous factors, including supply and demand dynamics, macroeconomic indicators, company-specific events, and investor sentiment. This makes the prediction of stock prices a complex problem that requires a multi-faceted approach (Malkiel, 2003).

One of the primary theories used in financial modeling is the Efficient Market Hypothesis (EMH), which suggests that all available information is already reflected in stock prices, making it impossible to consistently achieve above-average returns (Fama, 1970). However, machine learning models challenge this hypothesis by identifying patterns and signals within large datasets that may not be immediately apparent to human analysts (Lo, 2004).

### Time-series models

Time-series analysis is a crucial aspect of stock price prediction. Traditional time-series models, such as Autoregressive Integrated Moving Average (ARIMA), have been used to model stock price movements (Box & Jenkins, 1976). However, these models often struggle to capture the non-linear relationships and complex dependencies present in financial data. Deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, offer a more powerful approach to capturing these intricate patterns (Goodfellow et al., 2016).

### Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM)

CNNs are typically used for image processing but can be adapted for time-series data to identify local patterns and trends (LeCun et al., 1998). In this project, CNNs are used to extract features from time-series data, which are then fed into LSTMs. LSTMs are a type of Recurrent Neural Network (RNN) designed to handle long-term dependencies, making them well-suited for modeling the sequential nature of stock price data (Hochreiter & Schmidhuber, 1997).

### Sentiment analys

Sentiment analysis is another important component of this project. Investor sentiment, derived from news articles and social media, can significantly impact stock prices (Bollen et al., 2011). By using Natural Language Processing (NLP) models like FinBERT, we can quantify the sentiment associated with company news and incorporate it into the model as an additional feature, providing a more comprehensive view of the factors influencing stock prices (Yang et al., 2020).

### The hybrid

The hybrid CNN-LSTM model combines the strengths of both architectures: CNNs for capturing local temporal patterns and LSTMs for modeling long-term dependencies (Zhang et al., 2019). This approach allows the model to learn complex relationships within the data, ultimately improving the accuracy of stock price predictions. Hyperparameter optimization techniques, such as Grid Search and Bayesian Optimization, are used to fine-tune the model and achieve the best possible performance (Bergstra & Bengio, 2012).

Overall, the theory behind this project is based on leveraging advanced machine learning techniques to identify patterns and signals in complex financial datasets. By integrating multiple data sources and employing a hybrid modeling approach, the project aims to provide a robust solution for predicting future stock prices.

# Metod

The methodology for this project involves a systematic approach that includes data collection, feature engineering, sentiment analysis, model training, evaluation, and backtesting. Each of these steps is crucial for building a reliable and accurate AI model for stock price prediction.

## Data Collection

Data collection is the foundational step of this project. It involves gathering data from various sources, including:

* **Stock Price Data**: Daily stock prices are obtained from APIs such as Alpha Vantage and Polygon.io and historical stock prices are fetched from a downloaded txt file.
* **Technical Indicators**: Indicators such as RSI, MACD, SMA, and Bollinger Bands are calculated using stock price data with the Ta-Lib library.
* **Macroeconomic Data**: Macroeconomic indicators, such as GDP and inflation rates, are collected from sources like St. Louis fed.
* **Sentiment Data**: News articles and social media posts from Yahoo, Twitter and Reddit are analyzed to capture investor sentiment using the FinBERT model.

During data collection, special care is taken to handle missing data. For example, market holidays or weekends can result in missing stock price values, which are addressed using interpolation or forward-filling techniques to ensure data consistency.

## Feature Engineering

Feature engineering involves transforming raw data into features that can improve the model's predictive performance. This includes:

* **Lag Features**: Creating lag features that capture historical stock prices to allow the model to learn from past trends.
* **Technical Indicators**: Integrating various technical indicators that are commonly used by traders to predict market movements.
* **Time-Based Features**: Including features such as day of the week and month to capture periodic trends in stock prices.

The data is then standardized to ensure that features are on a similar scale, which helps in improving the convergence of machine learning algorithms.

## Sentiment Analysis

Sentiment analysis is conducted to quantify investor sentiment from news articles and social media. The FinBERT model, a specialized version of BERT for financial data, is used to determine whether the sentiment is positive, negative, or neutral. The sentiment scores are then integrated into the dataset as additional features.

## Model Training

The model training phase involves building and training a hybrid CNN-LSTM model. The CNN component is used to capture local patterns in time-series data, while the LSTM component captures long-term dependencies.

* **Hybrid Model**: The model architecture consists of a CNN layer followed by LSTM layers. The CNN extracts spatial features from the time-series data, which are then fed into LSTM layers for temporal pattern learning.
* **Hyperparameter Optimization**: Techniques such as Grid Search and Bayesian Optimization are employed to find the best set of hyperparameters, including learning rate, batch size, and the number of LSTM units.

## Evaluation

The model is evaluated using metrics such as Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and the Sharpe Ratio. These metrics provide insights into the accuracy of the model's predictions and its performance in terms of risk-adjusted returns

## Backtesting

Backtesting is performed to validate the model's performance using historical data. This step helps in assessing how the model would have performed in real trading scenarios, providing a measure of its practical applicability.

# Resultat och Diskussion

The results of the project are currently limited due to the incomplete training of the model. The model has not been fully trained primarily due to constraints in compute power and the time required for training a hybrid CNN-LSTM model on the collected dataset. Despite this limitation, several insights have been gained during the development process.

The data collection and feature engineering phases were successful in creating a rich dataset that includes historical stock prices, technical indicators, macroeconomic data, and sentiment scores. The sentiment analysis conducted using the FinBERT model provided valuable insights into the impact of investor sentiment on stock price movements, which will be crucial for improving model accuracy in future iterations.

Although the model training is incomplete, preliminary results indicate that the integration of CNN and LSTM architectures has the potential to capture both local and long-term patterns in stock price data. The CNN component successfully extracts spatial features, while the LSTM component is capable of learning temporal dependencies, suggesting that the hybrid approach is a promising method for stock price prediction.

Challenges encountered during the project include the computational resources required for training deep learning models and the time needed for hyperparameter optimization. Future work will involve completing the model training, fine-tuning hyperparameters, and conducting a comprehensive evaluation and backtesting of the model.

In conclusion, while the final trained model is not yet available, the methodology and preliminary steps taken in this project demonstrate a promising approach to stock price prediction using machine learning. With additional computational resources and time, the model has the potential to provide accurate and reliable predictions that can be used in real-world trading scenarios.

# Conclusion

This project set out to build an AI model capable of predicting future stock prices by integrating data from multiple sources, including stock prices, macroeconomic indicators, and investor sentiment. The goal was to leverage a hybrid CNN-LSTM model to capture both local and temporal patterns in the data. Due to the large volume of data, the focus was narrowed to Nasdaq tech stocks, specifically 'AAPL', 'MSFT', 'GOOGL', 'TSLA', 'NVDA', 'META', 'AMZN', 'NFLX'. While the model has not been fully trained due to computational constraints, several important conclusions can be drawn based on the progress made.

Firstly, integrating diverse data sources, such as historical stock prices, technical indicators, and sentiment analysis, is critical to improving the predictive accuracy of stock price models. The feature engineering process ensured that relevant information was included in the dataset, allowing the model to learn from both financial metrics and investor sentiment. The use of FinBERT for sentiment analysis provided valuable insights into how news and social media can influence stock prices, highlighting the importance of qualitative data in stock price prediction.

Secondly, the hybrid CNN-LSTM model architecture shows promise in capturing complex patterns in stock price data. The CNN component is effective in extracting local features, while the LSTM component is well-suited for modeling temporal dependencies. This combination has the potential to outperform traditional time-series models that struggle with non-linear relationships and long-term dependencies.

Challenges faced during the project, such as the need for extensive computational power and time for model training, highlight the practical difficulties associated with developing deep learning models for financial applications. These limitations prevented the complete training and evaluation of the model; however, the preliminary steps taken demonstrate that the approach is feasible and can yield meaningful insights.

In conclusion, this project demonstrates that a multi-faceted approach, combining technical, macroeconomic, and sentiment data with advanced machine learning models, can enhance stock price prediction. Future work will focus on completing the model training, optimizing its performance, and conducting a comprehensive evaluation to assess its real-world applicability. With adequate resources, the hybrid CNN-LSTM model has the potential to provide reliable predictions, supporting investors in making informed decisions in a highly volatile market

# Reflections

Reflecting on this project, there are several key lessons learned and areas for improvement that stand out. One of the main reflections is the importance of computational resources when dealing with deep learning models for financial predictions. The project encountered significant challenges in model training due to limited computational power and the extensive time required to train a hybrid CNN-LSTM model on a large dataset

Another important reflection is the value of integrating multiple data sources to improve prediction accuracy. The inclusion of stock prices, technical indicators, macroeconomic data, and sentiment analysis provided a more comprehensive understanding of the factors influencing stock prices. However, managing and processing such diverse datasets also added complexity to the project, highlighting the need for efficient data handling techniques and advanced feature engineering to extract meaningful insights from the data.

The use of sentiment analysis with FinBERT was particularly insightful, as it demonstrated the significant impact that investor sentiment can have on stock prices. This approach could be further expanded by incorporating additional sources of sentiment, such as earnings call transcripts or analyst reports, to gain an even deeper understanding of market sentiment and its effects on stock prices.

In conclusion, while the project faced challenges related to computational limitations and incomplete model training, the overall approach and methodology provide a solid foundation for future work. By addressing the identified challenges and incorporating more powerful computational resources, the project could achieve its full potential in accurately predicting stock prices. The lessons learned from this project will be valuable for future endeavors in financial forecasting using machine learning.

# Appendix A

En bild som visar text, skärmbild, Graf, diagram

Automatiskt genererad beskrivning

En bild som visar skärmbild, text, Graf, diagram

Automatiskt genererad beskrivning

# References

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13, 281-305.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. Journal of Finance, 25(2), 383-417.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Lo, A. W. (2004). The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. Journal of Portfolio Management, 30, 15-29.

Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics. Journal of Economic Perspectives, 17(1), 59-82.

Yang, Z., Xu, W., Shen, D., & Yang, J. (2020). FinBERT: A Pretrained Language Model for Financial Communications. arXiv preprint arXiv:2006.08097.

Zhang, Y., Aggarwal, C., & Qi, G. (2019). Stock Price Prediction via Discovering Multi-Frequency Trading Patterns. Proceedings of the 28th International Joint Conference on Artificial Intelligence, 3622-3628.