Deep Learning and Sentiment Analysis for Stock Price Forecasting

Amber Lee, Wafer Hsu, Cindy Zhang, Hejia Zhang, Oretha Domfeh

Data Science for All/Women, Correlation One – August 2023

Introduction

- Novel machine learning tools and data availability can be used to study financial market trends.
- We forecast stock prices with a deep learning model called Long Short Term Memory networks (LSTM).
- Qualitative information is incorporated through sentiment analysis of news headlines.

Data

Stock prices. We analyze the top five S&P 500 companies by weight: Apple, Amazon, Google, Microsoft, and Nvidia. The time frame is January 10, 2018 to June 30, 2023. The target is the daily adjusted closing price, which we refer to as the stock price. We access the data through `Quandl` [1].

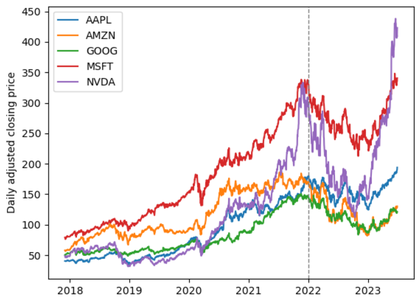

Figure 1:Daily stock prices

News Headlines. For every two-week period in our time frame, we conduct a keyword search of the five companies on Google News and retrieve the top 100 articles. The scraping was done with the `GNews` package [2].

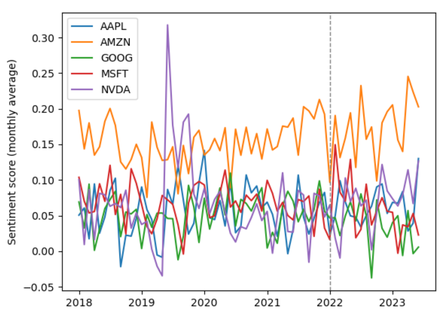

Figure 2:Monthly average sentiment score

Train/test split. We train our models from January 10, 2018 to December 31, 2021; then, we test from January 1, 2022 to June 30, 2023. The gray dashed lines in Figures 1 and 2 denote the split.

Methods: Sentiment Analysis

- NLP transforms large-scaled unstructured text data into structured and quantitative measurements of the sentimental opinions expressed by the text.
- Utilized VADER model from NLTK package [3]
- Sentiment analysis introduces positive, neutral, negative, compound.
- Included additional quantitative covariate: News Volume (the number of articles per day).
- For days without any news articles, imputed with `positive` = `negative` = 0, `neutral` = 1, and `volume` = 0.

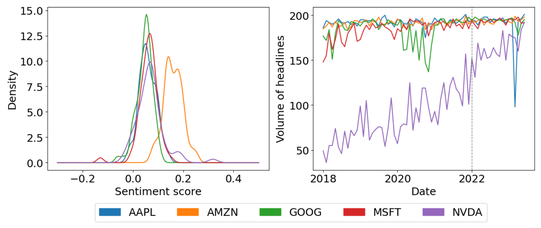

Figure 3:Sentiment Scores and News Volume Distribution

Methods: LSTM

What are LSTM's?

- LSTM networks are a type of recurrent neural networks, which contain cycles within the network activations to store contextual information. They take a sequence as input, making them appropriate for time series data.
- A LSTM unit is comprised of a cell and three gates for input, forgetting, and output. The cell is for memory, and the three gates regulates the memory.

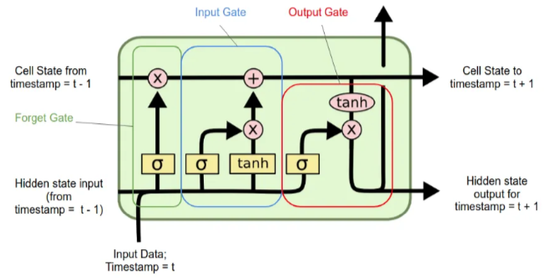

Figure 4:LSTM unit diagram from [4]

Implementation

1. Impute missing values with previous available stock price.
2. Choose parameters. From partial autocorrelation plots (Figure 5), a look-back window of 5 days was selected. We used two LSTM layers, each with 100 units.
3. Fit the model in Keras [5]. The baseline model is univariate – the input is the previous 5 days' stock prices. The baseline + sentiment model is multivariate that also takes daily sentiment score and headline volume as input.
4. Evaluate model on test data.

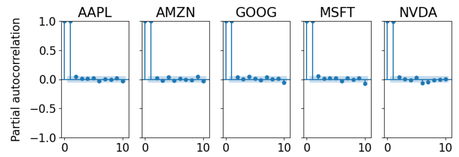

Figure 5:Partial auto-correlation of stock prices

Results

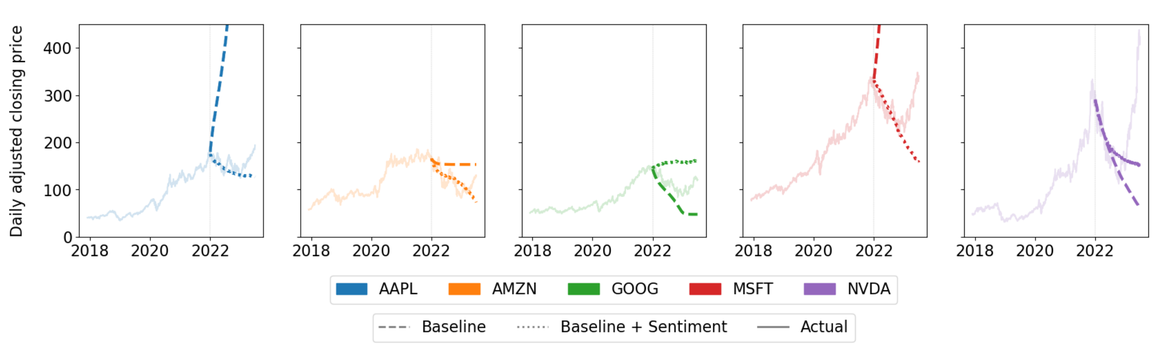

Figure 6:Figure caption

Conclusion

- Used web-scraping and performed sentiment analysis to transform large qualitative news information into quantitative measurement
- Implemented multivariate LSTM to tackle massive sequential stock prices prediction
- Our results indicated that by including qualitative news data as an input could significantly improve the prediction of stock price trends [Table 1]

	Baseline		Proposed	
Company	RMSE	MAE	RMSE	MAE
AAPL	765.8	604.1	23.2	17.5
AMZN	41.0	35.8	19.6	15.5
GOOG	39.6	33.9	47.8	44.3
MSFT	5776.3	3632.4	68.3	46.9
NVDA	122.3	76.8	85.0	54.7

Table 1:Prediction Performances of LSTM Models

Additional Information

Missing values. The news are published continuously while the closing prices are gathered on trading days. Thus, we need to handle the missing values.

Figure 7:Missing Values Distribution

References

[1] Quandl. Quandl python client. https://github.com/quandl/quandl-python, 2019.

[2] Muhammad Abdullah. Gnews. https://github.com/ranahaani/GNews, 2021.

[3] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[4] Ryan T. J. J. Lstms explained: A complete, technically accurate, conceptual guide with keras, Sep 2021.

[5] François Chollet et al. Keras. https://keras.io, 2015.