Stock Prediction using Long-Short Term Memory

Finn Zhan Chen

4th Year Project Report
Artificial Intelligence
School of Informatics
University of Edinburgh

2019

Abstract

The stock prediction tool can predict stock price of a public company one or more consecutive days using univariate and multivariate time series of technical indicators such as stock price, volume, and other more complex statistical methods.

Acknowledgements

Acknowledgements go here. Special thanks to project supervisor, Shay Cohen, for supporting the changing of the original project idea from Twitter sentiment analysis to stock prediction using technical data. Many relevant Python 3 libraries and online tutorials have been used to accelerate the development of the project. (To be completed)

Table of Contents

Chapter 1

Introduction

This Introduction chapter should include a clear and concise summary of your contributions (examples: adapting a suite of existing code; interpreting a theoretical algorithm; coding; testing; conducting an experiment) preferably as a bulleted list. TODO: To be completed

1.1 Problem

Stock prediction is hard. Sometimes simplicity might work better. Not everyone has or can afford automated trading. With the rising amount of different sort of data, this project aims to produce a user friendly stock prediction tool for young users who would like to invest by themselves with the support of a stock prediction tool.

1.2 Objective

This project focuses on the practical side of building a stock prediction tool purely based on technical analysis and technical indicators by leveraging the authors overall learning at the University of Edinburgh. The ideal end result is a functioning tool that assists the author in deciding the buy and sell moments in his stock investment portfolio, thus generating additional profit from the investment that would otherwise have not been possible. Optionally, a third party trading platform, such Degiro [15], in which the author uses to manually execute buy and sell stocks, could be automated using [7].

1.3 Tools

Various tools have been used during the development of the stock prediction tool in Python 3.

- Github repository has been set up for version controls
- Jupyter Notebooks have been used to speed up the entire development and debugging of the stock prediction tool
- Pycharm is also used to speed up development and debugging of object oriented designs on the main code of Jupyter Notebooks
- Guides
- Libraries (Github)
- Tensorflow

Chapter 2

Background

Historically, there are three main source of data that analysts or investors have used to predict the future stock price or the value of a companys stock price. Those are public technical analysis [1], fundamental analysis [12], and sentiment [11].

2.1 Technical Analysis

Technical analysis is an analysis methodology for forecasting the direction of prices through the study of past market data, primarily price and volume [6].

In technical analysis, technical indicators calculated from stock price sequence are used to predict the trend of future price changes. Many statistical methods have been proposed, but the results are insufficient in prediction accuracy [9].

Additionally, there are many other technical indicators beside price and volume. A technical indicator is a mathematical calculation based on historic price, volume, or open interest information that aims to forecast financial market direction [10].

2.2 Fundamental Analysis

Fundamental analysis is a method of evaluating a security in an attempt to assess its intrinsic value, by examining related economic, financial, and other qualitative and quantitative factors.

Fundamental analysis includes:

- Macroeconomic factors
  - Economy
  - Industry
  - Market

- Microeconomic factors
  - Company's financial reports
  - Company's key stats
  - Company's management

The end goal of fundamental analysis is to produce a quantitative value that an investor can compare with a security's current price, thus indicating whether the security is undervalued or overvalued.

A large body of evidence demonstrates that ratios of measures of fundamental value to market value systematically predict future stock returns. These ratios cmpare estimates of "intrinsic" values based on accounting data to observed market prices. They range from simple ratios such as earnings-to-price and book-to-market to ratios based on more sophisticated valuation models [3].

2.3 Sentiment Analysis

Sentiment Analysis also known as Opinion Mining is a field within Natural Language Processing that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression such as the polarity (positive or negative) towards a topic. (TODO: More on news)

Sentiment is the main and most subjective source of information that people expresses on news and social media that could be used to predict the stock price. [8] and [2] have successfully predicted the correlation of tweets and the stock market index such as Down Jones Industrial Average (DJIA).

However, news and sentiment analysis are often restricted by platforms (Twitter, US News Paper, etc...) and the diversity of the people who uses the platform. For example, [8] noted that the research could not capture the real public sentiment, because it only considers the twitter sentiment of english speaking people. Its possible to obtain a higher correlation if the actual mood is studied. It may be hypothesized that peoples mood indeed affect their investment decisions, hence the correlation. But in that case, theres no direct correlation between the people who invest in stocks and who use twitter more frequently, though there certainly is an indirect correlation.

2.4 Literature Review

Some papers focuses on generic indexes and not company specific stock. [TODO: To be completed]

2.4.1 Models for stock prediction

2.4.2 Univariate Time Series

Univariate time series: Only one variable is varying over time

2.4.3 Multivariate Time Series

Multivariate time series: Multiple variables are varying over time.

2.4.4 Long Short Term Memory

Chapter 3

Data Collection

In this chapter, data sources from technical analysis, fundamental analysis, and sentiment analysis for stock price prediction are explored, experimented, and evaluated. Below data sources are mostly US companies and are in dollars.

3.1 Technical Analysis

Many freemium (free without limitation), pseudo-freemium (limited features under the freemium version), and premium (requires fees from $500+) data sources of technical analysis are available.

Freemium data source:

- Alpha Vantage [16] provides:
  - stock price data from x until now in the US equities
  - 52 additional technical indicators available from X until now in the US equities
  - More to be udded here TODO
- IEX [4] provides:
  - stock price data for the past 30 days
  - volume data for the past 30 days

Pseudo-freemium data source:

- SimFin [14] provides:
  - 2000 API calls per day for free
  - daily adjusted stock price data from 1st January 2007 until now in the US

Premium data source:

- Zacks [13] provides:
  - no technical indicators
- Intrinio [5] provides:
  - "intraday and historical stock prices"
- Xignite [17] provides:
  - "end of day stock data and historical stock prices in the US and international equities"

After a thorough evaluation, Alpha Vantage is concluded to provide the best data overall for technical analysis to be fed into the tool.

3.2 Fundamental Analysis

Many freemium (free without limitation), pseudo-freemium (limited features under the freemium version), and premium (requires fees from $500+) data sources of fundamental analysis are available.

Fundamental analysis includes:

- Macroeconomic factors
  - Economy
  - Industry
  - Market
- Microeconomic factors
  - Company's financial reports (which includes data such as revenue, gross profit, cost of revenue, operating revenue, net income, research and development, operating expense, current assets, total assets, total liabilities, current cash, current debt, total cash, total debt, shareholder equity, cash change, cash flow, and operating gains losses)
  - Company's key stats (which includes data such as marketcap, beta, week52high, week52low, week52change, shortInterest, shortDate, dividendRate, dividendYield, exDividendDate, latestEPS, latestEPSDate, sharesOutstanding, returnOnEquity, consensusEPS, numberOfEstimates, EBITDA, revenue, grossProfit, cash, debt, ttmEPS, revenuePerShare, revenuePerEmployee, peRatioHigh, peRatioLow, EPSSurpriseDollar, EPSSurprisePercent, returnOnAssets, returnOnCapital, profitMargin, priceToSales
  - Company's management (from text descriptions on company's public strategic report)

Data available for machine learning need to be concrete and quantitative (time series). Therefore, only microeconomic factors such as a company's financial reports and key stats area readily available to be exploited.

Freemium data source:

- Alpha Vantage [16] provides:
  - does not provide fundamental analysis
- IEX [4] provides:
  - only has fundamental data from company's financial report for the past 4 quarters in the US equities
  - the company seems to be negotiating to introduce a greater number of quarters available

Pseudo-freemium data source:

- SimFin [14] provides:
  - fundamental data from company's financial report for the past 26 quarters for a limited number of companies from 01/01/2009 to 23/08/2018 in the US equities

Premium data source:

- Zacks [13] provides:
  - ”stock price consensus data”
  - ”fundamental data from company's financial report”
  - ”analyst's estimates for companies”
  - ”analyst's ratings”
- Intrinio [5] provides:
  - ”standardized US company fundamentals”
  - ”hundreds of fundamental ratios & metrics”
- Xignite [17] provides:
  - ”fundamentals, financial statements and ratios on U.S. and global companies”

The weakness of the available quantitative data are that there are limited amount of data points of fundamental analysis because the data are only released quarterly (4 times a year).

Under preliminary investigation, some fundamental indicators have strong correlation with the share price as shown in Figure 3.1 and 3.2. However, due to the lack of data points (quarterly reports) it is restricted to quarterly stock prices

```
1.000 AMZN.57: Market Capitalisation
0.999 AMZN.58: Enterprise Value
0.995 AMZN.37: Depreciation & Amortisation
0.986 AMZN.36: Total Equity
0.976 AMZN.31: Share Capital
0.967 AMZN.33: Retained Earnings
0.949 AMZN.19: Net PP&E
0.936 AMZN.10: Interest expense, net
0.925 AMZN.22: Total Noncurrent Assets
0.915 AMZN.6: SG&A
0.913 AMZN.23: Total Assets
0.904 AMZN.9: EBITDA
0.892 AMZN.17: Receivables
0.884 AMZN.28: Total Noncurrent Liabilities
0.866 AMZN.29: Total Liabilities
0.568 AMZN.25: Accounts Payable
0.508 AMZN.16: Cash and Cash Equivalents
0.497 AMZN.46: Operating Margin
0.269 AMZN.50: Current Ratio
0.264 AMZN.51: Liabilities to Equity Ratio
0.235 AMZN.41: Cash From Investing Activities
0.196 AMZN.53: EV / EBITDA
0.129 AMZN.56: Operating Income / EV
-0.033 AMZN.43: Net Change in Cash
-0.119 AMZN.39: Cash From Operating Activities
-0.183 AMZN.42: Cash From Financing Activities
```

Figure 3.1: Correlation between stock price and various fundamental indicators in graphs

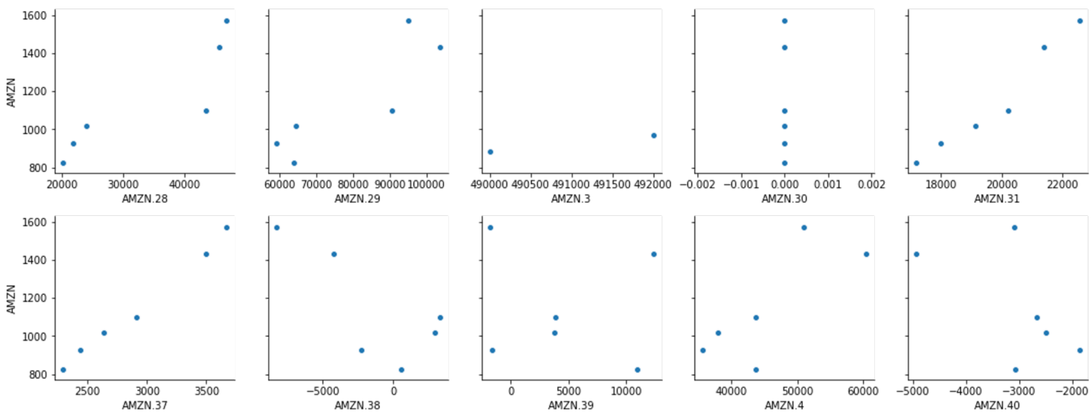

Figure 3.2: Correlation between stock price and various fundamental indicators in values

3.3 Sentiment Analysis

After research, the main source of sentiment available are news and social media data. News are harder to analyse and a decent API for historical news retrieval has not been discovered. On the other hand, Twitter has been experimented in the past for stock market prediction [8] [2] [18].

Pseudo-freemium data source:

- Twitter provides:
  - only history of 7 days can be retrieved from the Twitter's free API
  - other text processing challenges for social media opinions
  - tweets about limited number of public companies
  - sometimes multiple and ambiguous sentiments in tweets
  - not Tweets filtering to get the relevant tweets discussing a particular company
  - large amount of tweets (assume 2GB per day, 10 years require above 7.3TB) only if premium edition is obtained
  - some available sources of scraped twitter data but have been filtered and limited to [18] 01/01/2014 to 31/03/2016

Therefore, it is concluded that although there is potential to add sentiment analysis in the stock prediction it is not a very good data source due to many limitations.

3.4 Selection of data source

From the above preliminary research, it is concluded fundamental analysis and sentiment analysis are not practical given the fact that the data is sparse, hard to preprocess, and other limitations. Therefore, given the freely reliable available data for technical analysis from Alpha Vantage [16], Alpha Vantage is the main data source for predicting stock prices in the stock price prediction tool.

Chapter 4

System Design

Here discussed how the models are designed.

4.1 Object Oriented Programming

Baseline and LSTM models are written using the object oriented approach. This allows easier debugging and faster development of the prediction tool

4.2 Real Time

Data provided by Alpha Vantage is real time, hence making next day or multiple next days stock price prediction are possible. The code of the LSTM models are built such that it can conduct one-step or multiple-step forecasting for the stock price of any single public company in the US market when the LSTM models is trained and configured correctly.https://www.overleaf.com/project/5c48655efed02749045298b9

4.3 One-step Forecasting Model

This model allows the tool to make stock price prediction for the following day.

4.3.1 Baseline One-step Forecasting Model

This is the baseline model (persistence forecast model) cite that will be compared to the following machine learning models for evaluation. Additionally, other machine learning models discussed in the literature review will also be used to evaluate the stock prediction tool. 4.1 is an example.

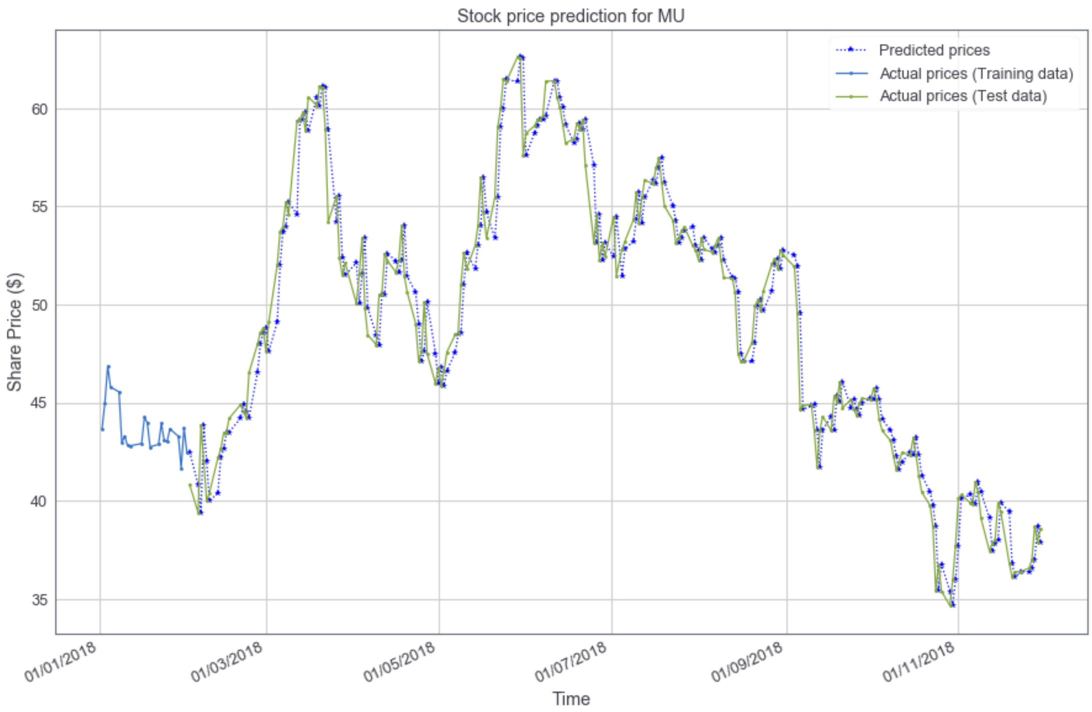

Figure 4.1: Baseline one-step persistence forecast model for MU from 01/01/2018 to 01/12/2018

4.3.2 Univariate One-step Forecasting LSTM Model

This model will use a single time-series in order to predict the stock price of next day. This time series is the historical stock price of a public company in the US. 4.2 and 4.3 are example.

4.3.3 Multivariate One-step Forecasting LSTM Model

This model will use multiple time series in order to predict the stock price of next day. This time series are combination of historical stock price and other technical indicators of a public company in the US.

TODO: to be completed

4.4 Multi-step Forecasting Model

This model allows the tool to make stock price prediction for more than one day ahead.

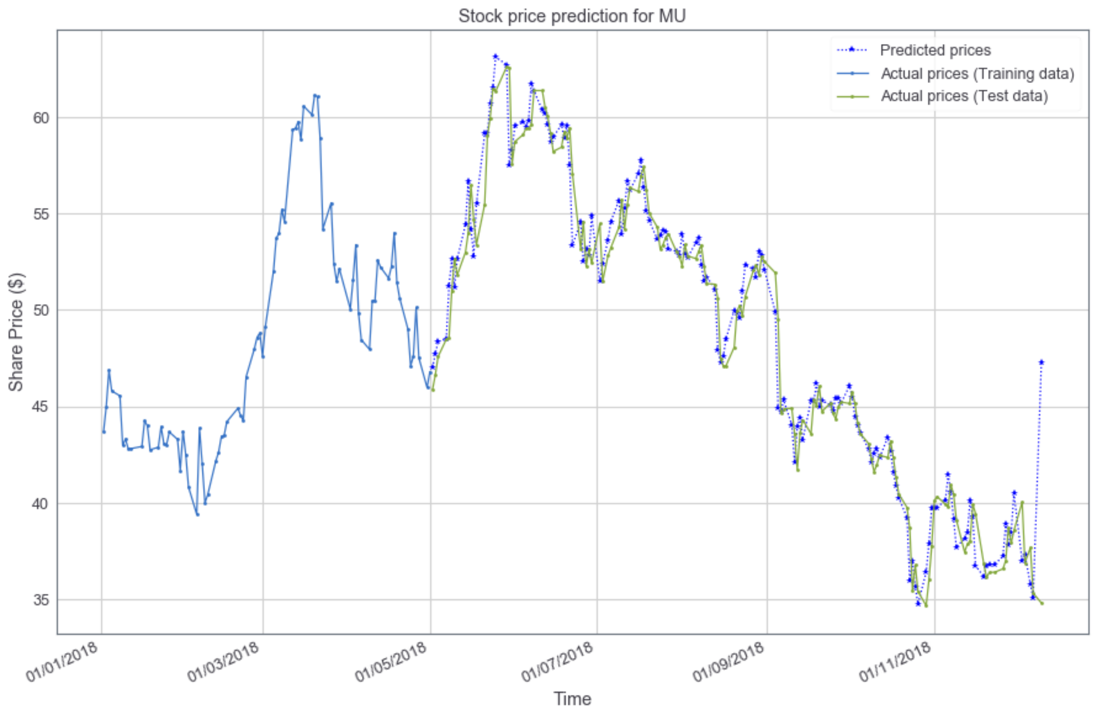

Figure 4.2: One-step LSTM forecast model for MU from 01/01/2018 to 01/12/2018

4.4.1 Baseline Multi-step Forecasting Model

This is the baseline model (persistence forecast model) cite that will be compared to the following machine learning models for evaluation. Additionally, other machine learning models discussed in the literature review will also be used to evaluate the stock prediction tool. 4.4 in an example.

4.4.2 Univariate Multi-step Forecasting Model

This model will use a single time-series in order to predict the stock price multiple days ahead. This time series is the historical stock price of a public company in the US.

TODO: to be completed, soon to be finished

4.4.3 Multivariate Multi-step Forecasting Model

This model will use multiple time series in order to predict the stock price multiple days ahead. This time series are combination of historical stock price and other technical indicators of a public company in the US.

TODO: to be completed

Figure 4.3: One day one-step LSTM forecast model for MU

4.5 More complex training data

Instead of just using one day of training data which all above models do, we could use multiple previous days of data to predict the future?

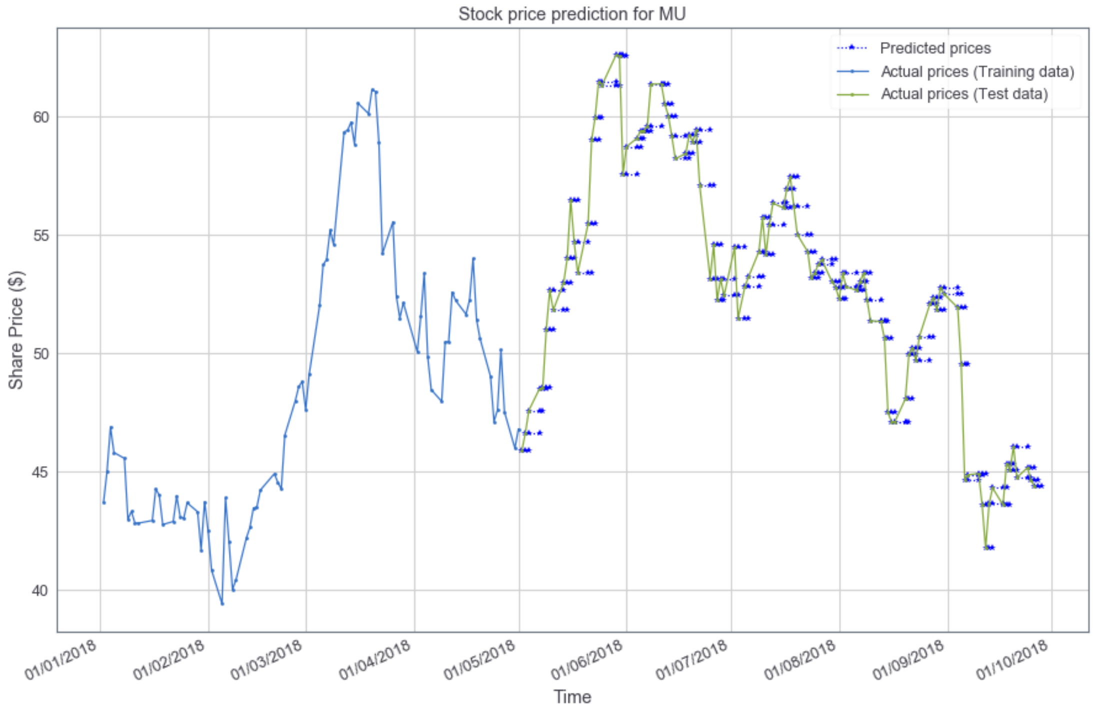

Figure 4.4: Baseline multi-step persistence forecast model for MU from 01/01/2018 to 01/12/2018

Chapter 5

Stock Price Prediction Implementation

TODO: To be completed. Expected to finish this by end of February. The rest of the time will be used to write up the report and ask for feedback.

5.1 Data Preprocessing for Supervised Learning in LSTM

5.2 Training

Already done, unsure on to explain it. Shall I copy paste code and explain?

5.3 Parameter Optimisation

TODO: To be completed

5.4 Prediction Visualisation

Following in an example

Chapter 6

Evaluation

6.1 Metrics

6.1.1 APE (Absolute Percentage Error)

TODO: To be completed.

6.1.2 AAE (Average Percentage Error)

. TODO: To be completed.

6.1.3 ARPE (Average Relative Percentage Error)

TODO: To be completed.

6.1.4 RMSE (Root Mean Squared Error)

Completed, see

6.1.5 MCC (Matthews Correlation Coefficient)

TODO: To be completed.

6.2 Obtaining Robust Result

TODO: To be completed.

6.3 One-step Forecasting Result

TODO: To be completed.

6.3.1 Univariate Time Series

TODO: To be completed.

6.3.2 Multivariate Time Series

TODO: To be completed.

6.4 Multi-step Forecast Result

TODO: To be completed.

6.4.1 Univariate Time Series

TODO: To be completed.

6.4.2 Multivariate Time Series

TODO: To be completed.

Chapter 7

Conclusions and Future Work

TODO: To be completed.

7.1 Fundamental and Sentiment Analysis

As demonstrated 3.2, it could be interesting to incorporate this new source of data on top of the technical data based on technical analysis for the stock market prediction tool

7.2 Automating Pseudo-Real Trading Platform

TODO: To be completed.

Bibliography

[1] George S Atsalakis and Kimon P Valavanis. Surveying stock market forecasting techniques–part ii: Soft computing methods. *Expert Systems with Applications*, 36(3):5932–5941, 2009.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[3] Patricia M Dechow, Amy P Hutton, Lisa Meulbroek, and Richard G Sloan. Short-sellers, fundamental analysis, and stock returns. *Journal of Financial Economics*, 61(1):77–106, 2001.

[4] IEX Group. The reliable, enterprise api that powers iex. free for everyone. https://iextrading.com/developer/docs/, 2019.

[5] Intrinio. Your financial markets data api. https://intrinio.com/, 2019.

[6] Charles D Kirkpatrick II and Julie A Dahlquist. *Technical analysis: the complete resource for financial market technicians*. FT press, 2010.

[7] Pedro Ladaria. Unofficial node.js api client for degiro's trading platform. https://github.com/pladaria/degiro, 2019.

[8] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15, 2012.

[9] Hirotaka Mizuno, Michitaka Kosaka, Hiroshi Yajima, and Norihisa Komoda. Application of neural network to technical analysis of stock market prediction. *Studies in Informatic and control*, 7(3):111–120, 1998.

[10] John J Murphy. *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[11] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, 2015.

[12] Jane A Ou and Stephen H Penman. Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics*, 11(4):295–329, 1989.

[13] Zacks Investment Research. Stock research, analysis, and recommendations. http://www.zacksdata.com/, 2019.

[14] SimFin. Online research platform with free data deed and api. https://simfin.com/, 2019.

[15] Degiro UK. Stockbroking online, invest worldwide. https://www.degiro.co.uk/, 2019.

[16] Alpha Vantage. Realtime and historical stock data for free. https://www.alphavantage.co/documentation/, 2019.

[17] Xignite. Market data feed and api. http://www.xignite.com/, 2019.

[18] Yumo Xu and Shay B Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1970–1979, 2018.