Contents

[1 Introduction 2](#_Toc536012405)

[Problem 2](#_Toc536012406)

[Objective 3](#_Toc536012407)

[Tools 3](#_Toc536012408)

[2 Background 4](#_Toc536012409)

[Technical Analysis 4](#_Toc536012410)

[Fundamental Analysis 4](#_Toc536012411)

[Sentiment Analysis (Twitter) 4](#_Toc536012412)

[LSTM best model for stock prediction 5](#_Toc536012413)

[Previous Work / Literature Review 5](#_Toc536012414)

[3 Data Collection 6](#_Toc536012415)

[Time Series Data Collection (real-time, offline) 6](#_Toc536012416)

[Time Series Data Preprocessing 6](#_Toc536012417)

[4 System Design 6](#_Toc536012418)

[Object Oriented Programming 6](#_Toc536012419)

[One-step forecasting LSTM Model 6](#_Toc536012420)

[Baseline one-step forecasting model 6](#_Toc536012421)

[Univariate one-step forecasting LSTM 6](#_Toc536012422)

[Multivariate one-step forecasting LSTM 6](#_Toc536012423)

[Multi-step forecasting LSTM Model 6](#_Toc536012424)

[Baseline multi-step forecasting model 6](#_Toc536012425)

[Univariate multi-step forecasting LSTM 6](#_Toc536012426)

[Multivariate multi-step forecasting LSTM 6](#_Toc536012427)

[Realtime 6](#_Toc536012428)

[5 Stock price prediction Implementation 7](#_Toc536012429)

[Parameter optimisation 7](#_Toc536012430)

[Training 7](#_Toc536012431)

[Data Preprocessing for Supervised Learning in LSTM 7](#_Toc536012432)

[Correlation with technical indicators 7](#_Toc536012433)

[Correlation with fundamental indicators 7](#_Toc536012434)

[Prediction Visualisation 7](#_Toc536012435)

[Obtaining robust result 7](#_Toc536012436)

[One step Forecasting 7](#_Toc536012437)

[Univariate Time Series 7](#_Toc536012438)

[Multivariate Time Series 7](#_Toc536012439)

[Multi-step result 7](#_Toc536012440)

[Univariate Time Series 7](#_Toc536012441)

[Multivariate Time Series 7](#_Toc536012442)

[6 Evaluation 7](#_Toc536012443)

[Metrics 7](#_Toc536012444)

[APE (absolute percentage error) 7](#_Toc536012445)

[AAE (average percentage error) 7](#_Toc536012446)

[ARPE (average relative percentage error) 7](#_Toc536012447)

[RMSE (root mean squared error) 7](#_Toc536012448)

[MCC (Matthews correlation coefficient) 7](#_Toc536012449)

[One step Forecasting 7](#_Toc536012450)

[Univariate Time Series 7](#_Toc536012451)

[Multivariate Time Series 7](#_Toc536012452)

[Multi-step result 7](#_Toc536012453)

[Univariate Time Series 7](#_Toc536012454)

[Multivariate Time Series 7](#_Toc536012455)

[Comparing baseline result using various metric 7](#_Toc536012456)

[7 Conclusions and Future Work 8](#_Toc536012457)

[Fundamental analysis and sentiment considered 8](#_Toc536012458)

[Online self-learning model 8](#_Toc536012459)

[Attributions 8](#_Toc536012460)

[Tutorials 8](#_Toc536012461)

Title and Abstract

# 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.

## Problem

Stock prediction is hard. Sometimes simplicity works better. Maybe works well. Not everyone has automated trading especially

## Objective

This project focuses on the practical side in which could bring the author’s overall learning at the University and consists of building a stock prediction tool purely based on technical analysis and technical indicators in the hope of automating stock trading on platforms like Degiro. The stock prediction tool can predict stock price of a public company one or more consecutive days.

Technical analysis is a trading discipline employed to evaluate investments and identify trading opportunities by analyzing statistical trends gathered from trading activity, such as price movement and volume.  (<https://www.investopedia.com/terms/t/technicalanalysis.asp>)

Additionally, other technical indicators are also used. A technical indicator is a mathematical calculation based on historic price, volume, or open interest information that aims to forecast financial market direction. <https://en.wikipedia.org/wiki/Technical_indicator#cite_note-1>

52 technical indicators are available in Table A (<https://www.alphavantage.co/documentation>)



## Tools

Jupyter Notebook

Libraries?

Pycharm

Degiro

# 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 company’s stock price (give support/reference). Those are public sentiment, technical analysis, and fundamental analysis. In this project, technical analysis is the only source of data to predict stock price. This is justified below.

## Technical Analysis

In technical analysis, technical indexes calculated from 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. (<http://sceweb.uhcl.edu/boetticher/ML_DataMining/Kimoto.pdf> )

## 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 analysts study anything that can affect the security's value, including macroeconomic factors (e.g. economy and industry conditions) and microeconomic factors (e.g. financial conditions and company management). (<https://www.investopedia.com/terms/f/fundamentalanalysis.asp>)

Data available for machine learning are the quarterly company reports which includes performance factors such as revenues, earning per share, pricing to book ratio, etc…

## Sentiment Analysis (Twitter)

Sentiment is the main and most subjective source of information that users expresses on social media that impact the stock price.

* such to how many tweets from certain past days (only history of 7 days) can be retrieved from the fremium API.
* Limited number of public companies are discussed
* Other text processing challenges
  + Ambiguous sentiments in one tweets
* Many not relevant tweets
* Large sizes for large amount of tweets (around 1GB per day, 10 years require above 4TB)
* Although some other source of scraped twitter data is available at (StockNet datasets: <https://github.com/yumoxu/stocknet-dataset>) but are limited to 01/01/2014 to 31/03/2016.

More in depth discussion in the final report

A large body of evidence demonstrates that ratios of measures of fundamental value to market value systematically predict future stock returns. These ratios cömpare 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 (e.g., Farna and French, 1995; Lakonishok, Shleifer, and Vishny, 1994) to ratios based on more sophisticated valuation models such as Ohlson (1995) (e.g., Frankel and Lee, 1998; Dechow, Hutton, and Sloan, 1999). Given the well-documented predictive ability of these ratios with respect to future stock returns, they provide a natural starting point for investigating the trading strategies of short-sellers. (https://deepblue.lib.umich.edu/bitstream/handle/2027.42/35551/b2036010.0001.001.pdf?sequence=2&isAllowed=y)

The weakness of this is that

* There are limited amount of datapoints of fundamental analysis because the data are only released quarterly (4 times a year)
* SimFin API can only retrieve data of the past 26 quarters for a limited number of companies from 01/01/2009 to 23/08/2018
* IEX API only has data for the past 4 quarters. However, the company seems to be negotiating to introduce a greater number of quarters available.

## LSTM best model for stock prediction

## Previous Work / Literature Review

LSTM vs Feedforward network:

Feedforward network’s outputs are independent of each other thus cannot capture the dependencies

They might have optimised parameters

* <http://www.cs.cmu.edu/~bdhingra/papers/stock_hmm.pdf>
  + HMM
  + ARIMA (AutoRegressive Integrated Moving Average)
  + ANN (artificial neural network)
  + IBM, APPLE, TATA STEELS, DELL
* Deep Learning
* Convolutional NN
* Unsupervised
* Supervised
* Logistic regression
* Linear Regression 2) Stochastic Gradient Descent (SGD) 3) Support Vector Regression (SVR)L
  + S&P500 - <https://pdfs.semanticscholar.org/0096/f7f6b0724c5839163b0e851b12b32f8ec908.pdf>
* Radial basis function SVC:
* Polynomial SVC
* LSTM
  + <https://arxiv.org/pdf/1603.07893.pdf> Google
* Vector Auto Regression

# 3 Data Collection

## Time Series Data Collection (real-time, offline)

Technical Analysis

Fundamental Analysis

Sentiment Analysis

Freemium

* Alpha Vantage
  + [https://www.alphavantage.co/query?function=RSI&symbol=MSFT&interval=daily&time\_period=10&series\_type=open&apikey=**3OMS720IM6CRC3SV**](https://www.alphavantage.co/query?function=RSI&symbol=MSFT&interval=daily&time_period=10&series_type=open&apikey=3OMS720IM6CRC3SV)
  + **Github wrapper** [**https://github.com/RomelTorres/alpha\_vantage**](https://github.com/RomelTorres/alpha_vantage)
* IEX Trading
* SimFin

Premium

* Zacks (1000+$)
* Intrinio (200+$)
* XIgnite (1000+$)

## Time Series Data Preprocessing

Dates

No outliers due to reliability of the api

# 4 System Design

## Object Oriented Programming

## One-step forecasting LSTM Model

### Baseline one-step forecasting model

### Univariate one-step forecasting LSTM

### Multivariate one-step forecasting LSTM

## Multi-step forecasting LSTM Model

### Baseline multi-step forecasting model

### Univariate multi-step forecasting LSTM

### Multivariate multi-step forecasting LSTM

## Realtime

# 5 Stock price prediction Implementation

## Parameter optimisation

Hypothesis, more data = better, test it

## Training

## Data Preprocessing for Supervised Learning in LSTM

## Correlation with technical indicators

## Correlation with fundamental indicators

## Prediction Visualisation

## Obtaining robust result

Train 10 models and get the average of the result

## One step Forecasting

### Univariate Time Series

### Multivariate Time Series

## Multi-step result

### Univariate Time Series

### Multivariate Time Series

# 6 Evaluation

## Metrics

### APE (absolute percentage error)

### AAE (average percentage error)

### ARPE (average relative percentage error)

### RMSE (root mean squared error)

### MCC (Matthews correlation coefficient)

Training and Testing set

## One step Forecasting

### Univariate Time Series

### Multivariate Time Series

## Multi-step result

### Univariate Time Series

### Multivariate Time Series

## Comparing baseline result using various metric

* Univariate Persistence Model Forecast
* Multivariate Persistance Model Forecast
* Comparison with existing results from other papers (HMM, ARIMA, ANN)

# 7 Conclusions and Future Work

## Fundamental analysis and sentiment considered

The limitation of a purely technical analysis-based stock price prediction tool is that it misses many important things, and will

## Online self-learning model

[Alpha Vantage](https://www.alphavantage.co/documentation/)

[IEX](https://iextrading.com/developer/docs)

# Attributions

Holidays <https://pypi.org/project/holidays/> (we’re using American stocks)

Adding business days: <https://stackoverflow.com/questions/12691551/add-n-business-days-to-a-given-date-ignoring-holidays-and-weekends-in-python/23352801>

Alpha Vantage Wrapper: <https://github.com/RomelTorres/alpha_vantage>

Degiro: <https://github.com/pladaria/degiro>

Backbone for one-step forecasting LSTM <https://machinelearningmastery.com/time-series-forecasting-long-short-term-memory-network-python/>

Multi-step forecasting LSTM: <https://machinelearningmastery.com/multi-step-time-series-forecasting-long-short-term-memory-networks-python/>

# Tutorials

<https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/>

<https://www.quora.com/Which-regression-model-is-best-for-predicting-forecasting-stock-prices>