UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER THESIS:

FORECASTING OF STOCK PRICES BASED ON SENTIMENT ANALYSIS AND TIME SERIES ANALYSIS

Ljubljana, May 2022 MANDAL EMA

AUTHORSHIP STATEMENT

The undersigned Ema Mandal, a student at the University of Ljubljana, School of Economics and Business, (hereafter: SEB LU), author of this written final work of studies with the title "Forecasting of stock prices based on sentiment analysis and time series analysis", prepared under supervision of Professor Jurij Jaklič

DECLARE

1. this written final work of studies to be based on the results of my own research;
2. the printed form of this written final work of studies to be identical to its electronic form;
3. the text of this written final work of studies to be language-edited and technically in adherence with the SEB LU's Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the SEB LU's Technical Guidelines for Written Works;
4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;
5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the SEB LU in accordance with the relevant SEB LU Rules;
6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;
7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;
8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the SEB LU Study Information System;
9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;
10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, _____21.06.2022____ Author's signature: _____

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AAPL – Apple Inc.

ANN – Artificial Neural Network

AR – Autoregressive

ARIMA – Autoregressive Integrated Moving Average

BI – Business Intelligence

CRISP-DM - Cross Industry Standard Process for Data Mining

EMH – Efficient Market Hypothesis

EN – English

i.e. – In example

MA – Moving Average

NLTK – Natural Language Toolkit

TS – Time Series

TSA – Time Series Analysis

SA – Sentiment Analysis

SNS - Social Networking Services

USA – United States of America

VADER – Valence Aware Dictionary and Sentiment Reasoner

INTRODUCTION

While searching for ways to multiply their wealth, many people invest in something. Although highly risky, most profitable investments are stocks. Stock market, probably one of the most popular markets nowadays, is based on buying a stock – a share of a company (Voigt & O'Shea, n.d.) with the expectation that its price will increase in a certain period of time. This is why stock market predictions have become very popular, especially with the evolution of artificial intelligence and machine learning.

One could claim that markets are unpredictable. However, they are complex and chaotic systems with both a systemic and a random component. This is why using chaos theory, together with powerful algorithms can make a realistic stock market forecast, although it is precise only to a certain extent (I Know First, 2021).

Stock market prediction can be defined as a process of forecasting future value of the company stock. Prediction methodologies can be divided in three categories: Fundamental Analysis, Technical Analysis and Quantitative Technical Analysis. Fundamental analysis is based on the whole image of the company. It evaluates assets, profits and business trends of a company and based on that evaluation, it determines the value of its stocks. So, if the current price is lower than determined one, stock should be bought. On the other hand, technical analysis is based only on statistics generated by market activity (Hladik, n.d.). While technical analysis usually uses data from short period of time, quantitative technical analysis is focused on longer period and it is often used to evaluate financial stability of a company.

In the recent period, two approaches, both connected to technical analysis, have been very popular: Time Series Analysis, further referred as TSA, and Sentiment Analysis, further referred as SA. Both of them are constantly being researched and compared. Time Series in general is defined as set of observations taken in approximately equal time intervals, in order to predict future behaviour of observed object (Azhikannickel, 2019). This method is frequently used in predicting stock market because the data are highly time-variant (Alestair, V., Harpreet, T., Aquib, S., Prasenjit,B. & Ashish, R., 2016). This approach implies finding patterns in historical behaviour of value of the stock in order to approximate its future value. Various models have been developed using this approach and ARIMA model has shown great accuracy of 96.5% (Chauhan, 2020). Still, TSA is not perfect. Unpredictable events can always occur and can highly affect future prices of the stock.

Other approach, SA, is very popular in natural language processing area. It tries to predict the sentiment (e.g. emotion) from the text (Bharathi & Geetha, 2017). There are few different approaches to analyse sentiment. Subjective or objective classification classifies the text to be objective – based on facts, or subjective – includes feeling. Sentiment detection and sorting analyses subjective sentences, the ones that have emotions, and categorises them as

positive, negative or neutral. This approach then determines intensity of sentiment, from very negative to very positive. Third approach is lexicon-based analysis which calculates polarity of the observed text by comparing every work with predefined lexicon - dictionary of positive and negative words, with a positive or negative sentiment value assigned to each of the words. There is also Aspect-based SA which tries to analyse the sentiment of the sentence based on the combination of words and not by analysing each word for itself. (Kapoor, 2021)

In my thesis, I am using lexicon-based approach. For lexicon-based approaches, a sentiment is defined by its semantic orientation and the intensity of each word in the sentence. Since people are very responsive to news and opinions of influential people, it is reasonable to assume that news affect the global opinion towards some companies, which then influences value of their stocks. Because of this, this approach is becoming more and more popular in the field of stock market prediction. But, since Internet is full of malicious users, who don't affect much global sentiment, it is challenging to assure quality of this approach (Shah, Isah & Zulkernine, 2018). While first approach, TSA, observes longer period of time, the other one relies on the most recent news, so it is rational to assume that the combination of two can yield more accurate predictions.

The purpose of my master thesis is to compare the efficiency of two approaches for stock market prediction: market SA and TSA, as well as to investigate whether hybrid approach of the two would yield better results.

My goal is to explore which approach for stock market prediction, market SA or TSA, gives better results. I will also implement hybrid approach of the two and investigate if the accuracy of its prediction would be higher.

Research question of my thesis is: "Is there a significant difference in accuracy of combination of two stock market prediction approaches: market sentiment analysis and time series analysis in comparison with their individual uses?"

In order to collect enough information about the topic, first of all I collect secondary data from various scientific sources. Since stock market prediction is an interesting topic that is constantly being explored, I first have to present what has already been done and how can I elevate current results. I explore theoretically both sentiment and TSA approach and based on my research I try to infer in which cases it is better to use one or the other approach. After I collect enough theoretical knowledge about the topic, I apply it in practice. Both approaches are implemented on the company of my choice using RapidMiner – a software platform used for data preparation, deep learning, machine learning, predictive analysis and text mining and Python – high-level programming language. Both platforms are explained later in the thesis. All approaches are implemented according to CRISP-DM methodology (Cross-industry process for data mining). It is an open standard process model, commonly used among data mining professionals. This methodology includes six major phases:

Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment (Data Science Project Management, n.d.). In order to implement the approaches, I first get secondary data about historical stock prices of selected company and relevant tweets. One part of collected data is used for training the model, while the rest is used for testing. Based on the results I am able to make some conclusions about the efficiency of both approaches. In the end, I am implementing hybrid approach of the two on the same company in order to be able to conclude if it does more accurate prediction. I use appropriate tools to asses all three approaches and based on that to compare them.

My thesis is structured into four chapters. First is Theoretical background, which has four main subchapters. In the first one, I present the concept of Efficient Market Hypothesis (EMH). Then in Literature review of stock market prediction techniques, I explore what have others already written on this topic and what are their findings. Data Mining Project aims to define what a data mining project is and which are the different types of these kinds of projects. Lastly, in the Definition of relevant data mining and text mining techniques subchapter, I define and elaborate on the exact TSA and SA techniques used in my thesis in order to familiarise reader with them. Next chapter is Methodology and tools which is also divided into more subsections. In the first – Methodology CRISP-DM part, I introduce CRISP-DM, the methodology I use for modelling the approaches. Second one describes tools used – RapidMiner Studio and Python, and how exactly are they used in my research paper. Then I introduce Twitter – social network used for collecting data on which I will perform SA. In next subsection – Selected market, I introduce the company whose stock prices I am predicting. Last subsection present prototype of the hybrid model that is implemented later in the thesis. Third part, Implementation of suggested approaches, is divided on Data collection and preparation, where I write about how I collected and prepared the needed data and Modelling, where I explain technical aspects of implementation of all three approaches. In the fourth chapter - Evaluation, I analyse the obtained results and write about what I have concluded from the results, how are my conclusions connected to previous findings and were the goals set in the beginning met. I also present an answer to the research question of the thesis.

Later in Conclusion I look back on the whole paper and write my conclusions. I explain how my implementations made upgrade on the already collected findings. I also elaborate what were my limitations and/or errors and suggest on what to pay attention in further research.

1 THEORETICAL BACKGROUND

In the first chapter of my thesis, I first introduce the concept of Efficient market hypothesis (EMH), a hypothesis in financial economics, that suggests that share prices reflect all available information. Then, in order to reach the goals set and answer research question presented in previous part, I need to understand what has already been written on the proposed topic. Here, I first focus on the TSA and SA, so I can assess which models of both

approaches are better used on their own. Then I focus on how did others combine different models of both approaches, and what were the results. This will help me decide on which models I want to implement. This is all done in the first subchapter – Literature Review.

As both TSA and SA are data mining projects, I introduce what kind of project it is, how does it work, what are different types of data mining techniques, and which are the benefits of this kind of projects. Then, in Definition of Models, I introduce the models I later use in the implementation part. Since neither TSA nor SA are easy to understand or implement, I here define and thoroughly explain exact models that I use to implement both analyses, as well as their combination. For my thesis I decided to use ARIMA for TSA and Lexicon-based SA, which I develop on VADER Lexicon.

1.1 Efficient Market Hypothesis

Efficient market hypothesis (EMH), a hypothesis in financial economics, suggests that share prices reflect all available information. This means that market prices immediately react to new information which then implies that it is impossible to predict market fluctuations and purchase undervalued stocks. (Downey, n.d.)

EMH is derived from the concepts present by Eugene Fama in his book from 1970 "Efficient Capital Markets: A Review of Theory and Empirical Work". The basic idea of his research is that it is impossible to consistently "beat the market" – meaning to continuously have investment returns which outperform the overall market average. His theory suggests that in principle, an investor can be lucky and gain huge short-term profit by making an investment, but long-term he will not be able to always make profitable investments.

Fama's investment theory is based on a number of assumptions about securities markets and how they function. These assumptions cover the idea which is the essence of validity of the EMH - the belief that all information relevant to stock prices is freely and widely available, "universally shared" among all investors.

As there are always a large number of both buyers and sellers in the market, price movements always occur efficiently (i.e., in a timely, up-to-date manner). Thus, stocks are always trading at their current fair market value. The main conclusion of EMH is that since stocks always trade at their fair market value. If that is the case, then, in principle, it is not possible to buy stock for an undervalued price nor to sell it for an overvalued price. If that's true, then the only way investors can generate superior returns is by taking on much greater risk.

There are three variations of the hypothesis – the weak, semi-strong, and strong forms – which represent three different assumed levels of market efficiency.

1. Weak Form - The weak form of the EMH assumes that the prices of securities reflect all available public market information. However, there could be some new

information that is not yet publicly available as such it is not reflected in the price. It additionally assumes that past information regarding price, volume, and returns is independent of future prices. The weak form EMH implies that technical trading strategies cannot provide consistent excess returns because past price performance can't predict future price action that will be based on new information. Based on this from, it is not possible to predict future prices by doing technical analysis, still, fundamental analysis may provide a means of outperforming the overall market average return on investment.

2. Semi-strong Form - The semi-strong form of the theory disregards both technical and fundamental analysis. This form has the assumptions of weak form with additional assumption that prices adjust quickly to any new public information that becomes available. This new assumption makes fundamental analysis incapable to predict future price movements.
3. Strong Form - The strong form of the EMH holds that prices always reflect the entirety of both public and private information. This includes all publicly available information, both historical and new, or current, as well as insider information. Even information not publicly available to investors, such as private information known only to a company's CEO, is assumed to be always already factored into the company's current stock price. So, according to the strong form of the EMH, not even insider knowledge can give investors a predictive edge that will enable them to consistently generate returns that outperform the overall market average.

Both supporters and opponents of the EMH have strong arguments to support their views. Supporters of the EMH often argue their case based either on the basic logic of the theory or on a number of studies that have been done that seem to support it. A long-term study by Morningstar found that, over a 10-year span of time, the only types of actively managed funds that were able to outperform index funds even half of the time were U.S. small growth funds and emerging markets funds. Other studies have revealed that less than one in four of even the best-performing active fund managers proves capable of outperforming index funds on a consistent basis. Opponents of the EMH advance the simple fact that there ARE traders and investors – people such as John Templeton, Peter Lynch, and Paul Tudor Jones – who do consistently, generate returns on investment that outperform the overall market. According to the EMH, that should be impossible other than by being lucky. But, how can then same people "beat the market" over and over again? In addition, those who argue that the EMH theory is not a valid one point out that there are indeed times when excessive optimism or pessimism in the markets drives prices to trade at excessively high or low prices, clearly showing that securities, in fact, do not always trade at their fair market value.

Investors who believe in the validity of the EMH are more likely to invest in passive index funds that are designed to mirror the market's overall performance, and less likely to be willing to pay high fees for expert fund management when they don't expect even the best of fund managers to significantly outperform average market returns. On the other hand,

because research in support of the EMH has shown just how rare money managers who can consistently outperform the market; the few individuals who have developed such a skill are ever more sought after and respected. (Corporate Finance Institute, n.d.)

While weak form of EMH suggests that fundamental analysis might be able to predict market movement, semi-strong and strong form suggest that it is impossible to constantly have very accurate predictions and trade on prices that are not "fair". I do agree that it is unlikely that one could make high-profit investments 100% of the time, but I still believe that by implementing right analysis approach, one could predict price movement, and therefore make profitable investment, in majority of cases. In my thesis I do exactly that. I question EMH hypothesis by combining technical analysis, analysis based on historical stock prices, and fundamental analysis, analysing sentiment towards a market, to make price movement predictions that are accurate most of the time.

1.2 Literature review of stock market prediction techniques

Stock market prediction refers to a process of trying to predict future value of a company stock which is traded on exchange. The main goal of this process is profit. Predicting whether market will go up or down helps investors make their decision on whether and when to invest in certain company. According to EMH there is no analysis which can consistently produce risk-adjusted excess returns. While there are many research papers that support EMH, there is an equal amount of papers that challenge it. In this part of my thesis, I will review the papers that are contrary to EMH and which suggest that stock market can be predicted to some extent by using different analysis methods.

1.2.1 Time Series Analysis for stock market prediction

Many factors affect stock market and that is why its prediction is such a challenging task. However, lot of analysts believe that there are some patterns in behaviour of stock prices. That is where TSA comes in place. Stock market data highly vary on time. One of the most used methods of TSA in Stock Market Prediction are Artificial Neural Networks (ANN). ANNs have the potential to recognize hidden patterns in information which are highly important for predicting stock market (Alestair, Harpreet, Aquib, Prasenjit & Ashish, 2016). Even though this technique is not perfect, it can handle some drawbacks of other techniques like over-fitting, trapped in local minima and black box technique (Gandhmal & Kumar, 2019). According to Gandhmal & Kumar (2019), 29% of papers on the topic of Stock Market prediction use ANN.

Grigoryan (2015) made a case study on TAL1T stock of Nasdaq OMX Baltic stock exchange, trying to predict its behaviour using ANN-based model called NARX in combination with Principal Component Analysis (PCA) for feature extraction. In most of the time, their model predicted correct values with very low error results of MSE_{150} =

0.0011703 and MSE_{200} = 0.0034567. Here, MSE represents Mean Squared Error and is a common evaluation measure for prediction performances on 150 and 200 samples.

In their paper on prediction of Istanbul Stock Exchange market, Kara, Boyacioglu & Baykan (2011) concluded that average prediction performance of ANNs is significantly better than the one of the SVM (Support vector machines) model.

Further, Kim & Kang (2019) suggest that deep learning methods, with special attention to TSA have proven better in financial predictions than machine learning methods. In their paper, they compared multiple deep learning models and concluded that when it comes to sequential data which is more time-dependant, LSTM (Long short-term memory) produced the best results, especially with longer look-back days.

Another model used for TSA is ARIMA model, which stands for autoregressive integrated moving average. It is commonly used for linear problems, so in theory, high-precision results are not expected when using this model for stock price prediction. However, Adebiyi, Adewumi and Ayo (2014) examined the performance of this model on Dell stock index and their model had quite low forecast error, meaning that predicted prices were very close to actual prices. Mondal, Shit and Goswami (2014) have observed the accuracy of an ARIMA model on fifty six stocks from seven different sectors, and accuracy was higher than 85% for all seven sectors, reaching as high as 99% accuracy for certain stocks.

Idrees, Alam & Agarwal (2019) analysed Indian stock market and tested ARIMA model to make prediction of future monthly average stock prices. As they tested multiple (p, d, q) combinations, the (0, 1, 0) yielded best performance and had roughly a deviation of 5% mean percentage error. Even though it is not perfect, as it is impossible to predict future values with 100% accuracy, this model still performed significantly good as it could give and idea of how market will behave.

As we can see ARIMA model can be used even for long-term predictions, however it is more relevant for short-term decision making. Ariyo, Adewumi & Ayo (2014) have also used ARIMA model to predict Nokia Stock Index, as well as Zenith Bank Index. When testing their model, they had to use different (p, d, q) values for different stock. The reason for this is that not every market has same stationarity and as a consequence not the same number of differentiations can be done. Additionally, every market is unique and can't be generalized. This is why it is very important to adapt every model to the stock one is trying to predict. Their models for both Indexes showed very good results in short-term predictions, with standard error of regression being 3.58 and 0.787 for Nokia and Zenith Bank respectively.

1.2.2 Sentiment Analysis for stock market prediction

21st century is the era of influencers. As their name suggests, those people have a power to shape the opinions of some part of population. The main tools they use to spread their word

are social media and news. This is why it is believed that stock market behaviour is hugely affected by opinions of certain individuals. So in the recent time, SA has been widely researched topic in the area of stock market prediction. But, as we all know, social media posts are usually short with lot of misspellings and grammatically incorrect sentences. This is why researchers have divided opinion on whether or not SA can be used to accurately predict stock price movement.

As Nguyen, Shirai, & Velcin (2015) suggest in their research, not every model has the same accuracy for every stock. For example, in their research on 18 stocks, their Aspect-based Sentiment model which integrates sentiments in social media over one year for prediction of stock price movements, had average accuracy of 54,4%. But, the same model observed only on Amazon.com, Inc stocks shows 75% accuracy. Since, according to their paper, models which yield at least 56% accuracy are considered satisfying, the result for this particular stock is extremely good.

Indeed, SA has improved many previous stock market prediction techniques. In the paper of Bharathi & Geetha (2017), addition of SA to the moving average model, a technical analysis tool in which the actual index data is compared with its average taken over a period of time, has improved its accuracy from 64,32% to 78,75%.

Besides usual SA approach, which focuses on classification of sentiment of text, Ming, Wong, Liu & Chiang (2014) focus on creating a dictionary in which they pair news articles with stock prices. This approach helps in predicting the day's closing price using the articles of the day. The main positive side of this approach is that it avoids errors which can occur when evaluating sentiments as positive or negative.

Liew and Wang (2016) were comparing tweets analysis and IPO – Initial Public Offering performance and they concluded that there is a positive significant correlation between IPOs' average tweet sentiment and IPOs' first-day returns not only on the first trading day but also two or three days prior.

Abbes (2016) suggest retrieving tweets in a slightly different way than usual. In his work, he searched only for tweets containing $ symbol followed by the stock name. $ symbol makes sure that tweet is referring to financial subjects. Another interesting point he mentions is that positive and negative tweets don’t have the same intensity of impact on stock market behaviour. While implementing SA on tweets, instead of using already available SA tools like the Alien one from RapidMiner or TextBlob or VADER libraries from Python, he rather used couple of lexicons which had words classified as positive or negative. Each word in tweet was then compared with these lexicons and was given value 1 or -1 depending on whether it was positive or negative. All values were then summed up and depending on this score, tweet was given value 1, 0 or -1 meaning positive, neutral or negative. As his focus was on UK stock market, which has limited coverage on twitter results were not promising.

However, this is very interesting approach which could have a great potential in prediction of rise/fall of stock prices.

Smailovic, Grcar, Lavrac & Znidarsic (2013) used casuality analysis to compare daily change in number of positive tweets with daily change of stock prices. They concluded that for some companies that had significant changes in closing prices over observed period of time, sentiment of tweets was able to predict price movements. Even though the performance varies from company to company, it can still give strong indication if the price will rise or fall when share of positive tweets rises or falls significantly over 2 or more days.

In their research on the correlation of sentiment of tweets with prices for Bitcoin, Pano and Kashef (2020) emphasized the importance of cleaning the textual data before doing SA. The cleaned data had greater correlation with price movements than raw data. Another valuable conclusion from their research is that lower timespans, particularly one day, have better performance than longer time spans when doing simple SA with VADER.

1.2.3 Hybrid approach of Sentiment and Time Series Analysis for stock market prediction

All of the models mentioned above are effective to some extent, but each one can be improved. Using only one single model for prediction, although in some cases can show exceptional results, it can't be generalized. The future trend of stock market is not only related with historical prices, but also civil economy, financial policy, public opinion, global events and so on.

First, let's explain what a hybrid model actually is. Hybrid model, in the context of stock market prediction, represents combination of two or more forecasting techniques used together to make more accurate predictions. Many researchers have been investigating different combinations which could improve current models. According to the research taken by Koceska and Koceski (2014) where they reviewed various previously published financial time series modelling and prediction techniques, hybrid models make more accurate forecasts. The combination of various techniques can improve each-others drawbacks.

Tang, Yang and Zhou (2009) have tried to improve regular TSA algorithm by incorporating it with news text mining. The proposed model called NTF (News mining and Time series analysis based Forecasting), tested on Chinese stock data showed significantly lower average absolute difference rate in forecasted values than regular TSA algorithm used alone.

The hybrid model was studied on Chinese stock market also by Wang (2017). He extracted data from stock-related micro-blogs, and created 2 time series models. One was based on sentiment values and the other was based on stock index. He then created neural network model of index values as well as sentiment values which had better results than the one that

was fed only with index values. Interestingly, he suggests that if more external indicators were included, results would probably be even better.

Chou (2021) observed various time series models and their combination with SA models on 6 stocks from different sectors: Bank of America Corporation (BAC) in financial services, The Boeing Company (BA) in industrials, Exxon Mobil Corporation (XOM) in energy, Uber Technologies, Inc. (UBER) in technology, Johnson & Johnson (JNJ) in healthcare, and Apple Inc. (AAPL) in technology. Their model based on Historical Prices, Sentiment Scores, and Predicted Prices referred to as Model_8 showed best lowest errors for 5 out of 6 stocks.

Kedar and Kadam (2021) were observing SA approach and ARIMA algorithm for predicting future prices. While for stable markets, ARIMA had relatively high accuracy, for the volatile ones, analysing sentiment of tweets had a potential to improve the prediction accuracy.

In their research, Mehta, Malhar, and Shankarmani, (2021) were comparing different machine learning models to predict stock prices in not distant future. They conclude that LSTM are very good for long-term predictions, while ARIMA is better for short term predictions because it dynamically adjusts to changes. Additionally, they did twitter scraping and analysed the overall sentiment which helped them to adjust their predictions to be more accurate. Considering the fact that ARIMA is linear model, SA can help predicting sudden increases or decreases in the market prices.

Mohan, Mullapudi, Sammeta, Vijayvergia & Anastasiu (2019) suggest there is a strong correlation between news articles related to the company and its stock value. In their research paper they focused rather on finding correlation between daily closing prices and news articles from daily newspaper websites. They have tested different models on the same dataset. Some models used only historical prices as input while others combined historical prices with news analysis. They conclude that the RNN network that used combination of historical stock prices and daily polarity based on news texts had the lowest prediction error. This was especially notable on volatile markets that are not very stable over time.

"In the Indian stock market, stock costs are viewed as exceptionally fluctuating due to various factors such as political decision results, bits of gossip, budgetary news, public safety events and so on. This fluctuation behaviour makes it a difficult and challenging task to predict stock prices." (Kesavan, Karthiraman, Ebenezer & Adhithyan (2020). As such, this market is very good example to observe if SA can improve the prediction of stock movement. They proposed analysing sentiment of twitter data and combining it with historical prices. Their model predicted the price movement for the next day by comparing number of positive and negative tweets and prediction prices. It resulted with 3.05 percentage error which is very low error especially for this kind of market.

1.3 Data mining project

In my thesis I am analysing two types of data: historical prices and twitter data, as well as their combination in order to predict the market behaviour. As I use data to make a decision, this type of project is Data Mining Project. Therefore, in subchapters 1.2.1 Definition and process and 1.2.2 Types of Data Mining Techniques I explain what kind of project this is and different types of techniques that can be used for its implementation.

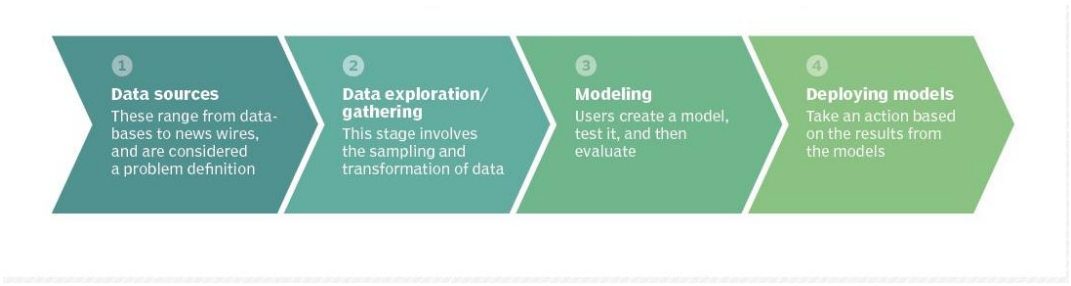
1.3.1 Definition and process

Data Mining is one of the key disciplines in data science and a crucial part of data analytics. It uses different analytics techniques in order to extract useful information from large data sets. It is a process of examining through those data sets with an aim to find patterns and/or relationships that can help to solve business problems. Based on that enterprises can predict future trends and make data-driven business decisions. The information generated with Data Mining process can be used in business intelligence (BI), analysis of historical as well as real-time data. Apart from business side, it has important role also in healthcare, government, scientific research, mathematics sports etc.

In my thesis the purpose of data mining process is to make informed decision on whether or not to invest in certain stock. First, I need to explain how does, in general, data mining process look like. The data mining process can be broken down into four main stages:

1. Data gathering – in the first stage of Data Mining project, one needs to identify relevant data which can be located in different source systems, a data warehouse or a data lake or it can be an external data source. In any case, all data needs to be moved to a one data lake which will be used in the remaining steps.
2. Data preparation – In many cases, raw data is not ready to be mined. This is why it has to be pre-processed and cleaned to make sure it is consistent and of high-quality.
3. Mining the data – after the data is prepared and appropriate data mining technique is chosen, one or more algorithms which do the mining are implemented. The algorithm usually has to be trained on sample dataset and is then implemented on the complete dataset.
4. Data analysis and interpretation – the final results are used to create analytical models and the findings are communicated, often through data visualization and the use of data storytelling techniques. (Stedman & Hughes, n.d.)

Figure 1: Four stages of data mining



Source: Stedman & Hughes (n.d.).

The process and its implementation in this research paper will be explained more in detail later in the Methodology part where I focus on the specific methodology used for my data mining project.

1.3.2 Types of data mining techniques

There are various techniques that can be used for data mining. One use case can be done with many different techniques, and they are all divided into two main types: Predictive Data Mining and Descriptive Data Mining. (Java T Point, n.d.)

Predictive Data Mining, as the name suggests, uses the data to make prediction on what might happen in the future. It can be divided into four subtypes:

- Classification Analysis – this type classifies elements of the dataset into different categories which are pre-defined as part of the data mining process. Some examples of classification method are k-nearest neighbour, logistic regression, decision trees etc.
- Regression Analysis – regression too is used to find relationships in datasets, by calculating predicted data values based on set of variables. Most common techniques are linear and multivariate regression.
- Time Series Analysis – Time Series is a sequence of data points recorded in, most often, regular time intervals. This analysis aims to predict future values based on those intervals.
- Prediction Analysis – This technique is usually used to predict the relationship that exists between both the independent and dependent variables as well as the independent variables alone.

Descriptive Data Mining tries to summarize the data into relevant information which can later be used for decision making. This type too can be divided into four subtypes:

- Clustering – this technique groups together elements that have some particular characteristics in common. Unlike Classification that collects the objects into predefined classes, clustering stores objects in classes that are defined by it.
- Summarization Analysis – summarization is used to group data in a compact way, for example calculating averages of a subset.
- Association rule mining – Association rules are if-then statements which spots relationships between different data elements. In order to assess those relationships, support and confidence criteria are used. Support counts how often the related elements appear in dataset, while confidence counts how many times an if-then statement is correct.
- Sequence Discovery Analysis – this type of data mining techniques looks for interesting patterns in data where set of events leads to another event. It is very similar to the TSA, however while TSA works with numerical data, Sequence Discovery Analysis usually works with discrete values.

In my thesis I am predicting future behaviour of stock market, therefore I use Predictive Data Mining. I implement TSA directly, while SA, in the way I implement it, is type of Prediction Analysis since it tries to find relationship between sentiment and stock price.

In general, the benefits of data mining come from the increased ability to uncover hidden patterns, trends, correlations and anomalies in datasets. With that information available, one can be improve business decision-making and strategic planning through a combination of conventional data analysis and predictive analytics. As mentioned before, I it is widely used across many industries (Stedman & Hughes n.d.).

1.4 Definition of relevant data mining and text mining techniques

As suggested in different literature sources, there are many different models for both TSA and SA. In general, it cannot be said that one model is better than other in all aspects, therefore it is important to recognise which model to use in order to answer your research question. After doing some preliminary research I have decided to use ARIMA model for TSA and Lexicon-based model for SA. Both of them are explained in this subchapter. I also explain the concept of hybrid model and it's purpose in my thesis.

1.4.1 ARIMA model for Time Series Analysis

One of the commonly used models for TSA in ARIMA - Autoregressive Integrated Moving Average. As every other model it has its advantages as well as disadvantages. However, it can be easily implemented especially using RapidMiner Studio. There are also many papers on which I can build upon and compare the results. So, let's explain what ARIMA stands for.

ARIMA is type of Univariate Time Series Forecasting. This means that it uses only information about past values of the time series to predict future values. ARIMA is an integration of two models: Auto Regressive (AR) and Moving Average (MA) model. So, in order to understand ARIMA itself, one first needs to understand concepts of these two models. (Prabhakaran, 2021).

Auto Regressive model is type of regression model where values of observed variable depends only on values in previous steps. The type of correlation is partial auto-correlation which means that it only considers direct impact of previous values on predicted value, but not the effect which previous values have on one another. The equation for AR model is formulated equation (1):

$$Y_t = \beta_1 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi_p Y_{t-p} \quad (1)$$

β is an intercept term estimated by model, Φ_1, Φ_2,…, Φ_p are weights of corresponding lagged values and are also estimated by model, while Y_{t-1}, Y_{t-2},…, Y_{t-p} represent lagged values. Value p is called lag order and it represents how many prior values are included in the model. We say that those values have a significant correlation with the predicted one (Rajbhoj, 2019).

While AR model considers impact of previous values on the predicted one, Moving Average model observes errors of previous forecasted values. It actually analyses how wrong were predictions in previous time periods in order to make better forecast for the current time-period. MA too uses partial auto-correlation. The equation (2) represents the formula of MA models.

$$Y_t = \beta_2 + \omega_1 \varepsilon_{t-1} + \omega_2 \varepsilon_{t-2} + \cdots + \omega_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

Similarly as in AR model, β is an intercept term estimated by model, ω_1, ω_2,…, ω_q are weights of corresponding lagged error values while ε_1, ε_2, …, ε_q are errors of previous forecasts and ε_t will be error of current forecast. Value q is the size of moving window which tells how many previous observation errors are directly impacting current observation (Rajbhoj, 2019).

Combination of the two models then takes into account both historical values and errors of previous forecast, and equation (3) is their combined equation.

$$Y_t = (\beta_1 + \beta_2) + \begin{matrix}(\Phi^1 Y_{t-1} + \Phi^2 Y_{t-2} + \cdots + \Phi_p Y_{t-p}) \\ +(\omega_1 \varepsilon_{t-1} + \omega_2 \varepsilon_{t-2} + \cdots + \varepsilon_{t-q} + \varepsilon_t)\end{matrix} \quad (3)$$

It is important to emphasize that both AR and MA model assume that the series is stationary, meaning that the mean and standard deviation of series are constant and that there is no seasonality. But, many time series do not satisfy this condition. In those cases data needs to be transformed. Commonly used transformation is difference of two steps in the time series.

In that case we are not observing Yt, but instead we are focusing on $Z_t = Y_{t+1} - Y_t$. If the stationarity condition is still not satisfied, then we apply new transformation where we observe $Q_t = Z_{t+1} - Z_t$. The number of transformations needed to make time series stationary is called order of differencing and is denoted by d. (Rajbhoj, 2019).

ARIMA model combines all above explained techniques. It first transforms the data and then applies AR and MA models. It is characterized by 3 main parameters: p, d and q explained previously.

1.4.2 Lexicon-based model for Sentiment Analysis

In order to implement SA, I decided to use lexicon-based model. In this model sentiment is defined by its semantic orientation and the intensity of each word in the sentence.

Lexicon represents the vocabulary of a person, language or branch of knowledge. So, in lexicon based sentiment analysis one needs to have a set of dictionary words which are already labelled as positive, negative or neutral. Besides that, these words usually have polarity , parts of speech and subjectivity classifiers, mood, modality etc. The model tokenizes the sentence, and each word is matched with the available words from dictionary to find out it's context and sentiment. The final score of sentence is usually based on average or sum of the scores of each word. (Roul, 2021)

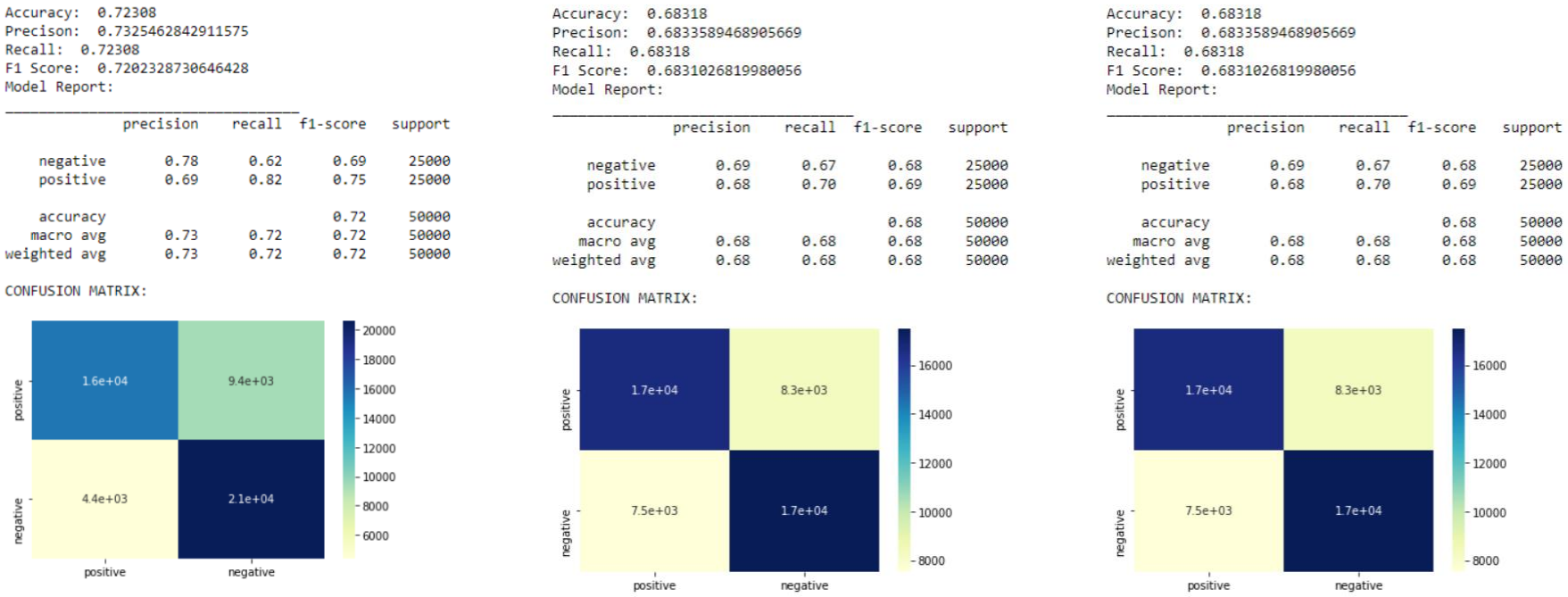
There are many available lexicons which can be used for this type of SA, and I will introduce few of them. First one is AFINN Lexicon. It is very simple and one of the most popular lexicons for SA. The latest version, AFINN-en-165.txt, contains 3382 words and their polarity scores. The words are manually rated with an integer between -5 and 5 by Finn Arup Nielsen. (Roul, 2021)

Another lexical resource for opinion mining is SentiWordNet Lexicon. It operates on the database provided by WordNet, which is composed of English words, grouped as synonyms into so-called synsets. Every synset is associated with three scores: positivity, negativity and objectivity (neutrality) score which have values between 0 and 1. (Sharma, 2021)

Third one is VADER, Valence Aware Dictionary and Sentiment Reasoner, lexicon and rule-based sentiment analysis tool which is specifically designed to analyse sentiments expressed in social media. It is available in NLTK Python package and can be applied directly to unlabelled text data. It can detect polarity as well as intensity of emotion in text. (Roul, 2021)

Roul (2021) has tested all three lexicons on same IMDB movie reviews dataset and results are shown in Figure 2 - AFFIN Lexicon, SentiWordNet and VADER form left to right respectively.

Figure 2: Performance of different lexicon-based sentiment models



Source: Roul (2021).

As we can see, for this dataset, AFFIN Lexicon and VADER had overall better performance than SentiWordNet. Driven by these results and the fact that in my thesis I use Twitter data, which is social media network, I decided to base my SA on VADER.

2 METHODOLOGY AND TOOLS

This chapter of the thesis explains which methodology I used in order to implement the proposed models for stock market prediction and answer the research question. As the process I implement is data mining process, I am using Cross Industry Standard Process for Data Mining, referred to as CRISP-DM. The process is composed of six phases: Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment, which are all explained in the first subchapter – Methodology CRISP-DM. In the first subchapter, I will explain the details of this methodology in detail

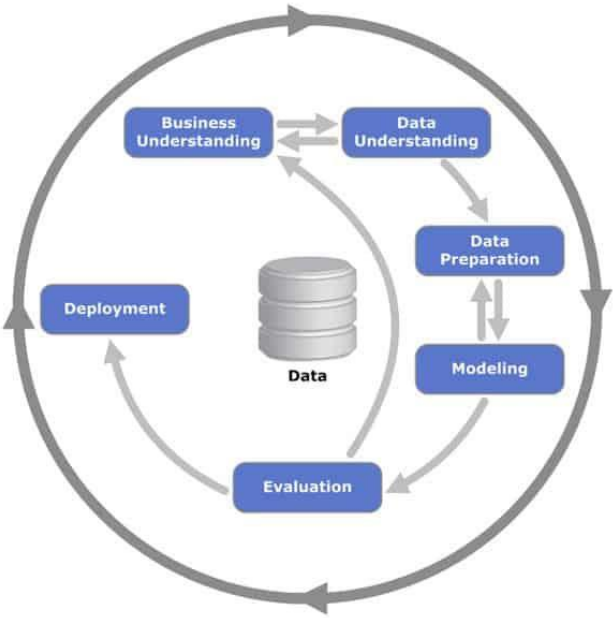
I implemented the models with RapidMiner Studio - software platform used for data preparation, machine learning, deep learning, text mining and predictive analysis, and Python - high-level programming language. Both softwares are very powerful and often used for implementation of different data mining processes. In subchapter 2.2 RapidMiner Studio and Python, I will introduce both of them and explain how I use them in my research. In subchapters 2.3 and 2.4 I introduce Twitter – social network I use to analyse sentiment towards market and I brief on background of company whose stock I a using to implement and evaluate the models.

2.1 Methodology CRISP-DM

As many research papers in the field of data mining, this one too is based on CRISP-DM methodology. CRISP-DM is widely used analytics model with six phases that describes the life cycle of a data mining project. Those six phases include:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modelling
5. Evaluation
6. Deployment

Figure 3: CRISP-DM Lifecycle



Source: Data Science Project Management (n.d.).

According to CRISP-DM, in order for a project to be successfully implemented, one first needs to understand the objectives and requirements of the project. Clear goals need to be defined, while having in mind available resources. There also needs to be defined a success factor, meaning that one needs to have scale which will measure a level of success of the project. And of course, in this phase, we need to make a clear plan of the whole process. This part of methodology is implemented in the Introduction as well as Theoretical background parts, where I defined research question, purpose and goals as well as explained the business problem. After having it all defined, second step would be collecting and understanding the data on which the project will be based on. This includes finding relevant data sources, documenting properties of the collected data, understanding the relationships among data and verifying its quality. I explain this in the following, Data collection part where I explain which data I used, and how did I collect it. The third step would be to clean the data. That means that I have to identify if there are some missing or incorrect values, which attributes are needed for my project or do I want to derive some new attributes from already existing ones. I also need to format the data and take any other steps that would improve the quality of data. This is usually the hardest and most time-consuming step, and it is explained later in the thesis. Having the data prepared, I then need a model that will use it. In practice, few models are selected and assessed. In this phase one also has to decide how much data will be used for training, testing and validation purposes. While this step concentrates more on the technical aspects and validity of models, the Evaluation step is more concerned about how models fit previously set business perspective. Here they are measured based on success factors defined in the first step. If none of the models is good enough, one needs to go one step back to improve them or restart the project from the

beginning. Otherwise, after one of the models is selected it needs to be deployed. Deployed project is then documented and reviewed (Data Science Project Management, n.d.).

2.2 RapidMiner Studio and Python

As noted previously, in order to implement the models for all SA, TSA and hybrid of the two, I used RapidMiner Studio and Python.

RapidMiner Studio is a powerful software platform used for data preparation, machine learning, deep learning, text mining and predictive analysis. It supports different steps of data mining project, from data preparation to validation and optimization. Each process in RapidMiner Studio consists of one or several operators that perform different tasks. It is a paid platform, but it has a free Educational Licence Program with access to all operators. It has large number of available operators and I will introduce the ones I used:

Retrieve - The Retrieve Operator loads a RapidMiner Object, which is often and ExampleSet, into the Process. Retrieving data this way also provides the meta data of the RapidMiner Object.

Sort - This operator sorts the data set provided at the input port. The complete data set is sorted according to a single or more attributes. The attributes to sort by are specified using the sort by parameter. For each attribute, sorting is done in ascending or descending order, depending on the setting of the sorting order parameter. The resulting data set is sorted by the first attribute, then subsets of the same value in the first attribute are sorted by the second attribute etc.

Generate Attributes - The Generate Attributes operator constructs new attributes from the attributes of the input ExampleSet and arbitrary constants using mathematical expressions. The attribute names of the input ExampleSet might be used as variables in the mathematical expressions for new attributes. During the application of this operator these expressions are evaluated on each example, these variables are then filled with the example's attribute values. Thus this operator not only creates new columns for new attributes, but also fills those columns with corresponding values of those attributes.

Lag - This operator performs a time series lag transformation on one or more attributes. Individual attributes can be lagged separately with different lag values by the parameter individual lags. In addition, a default lag for a set of attributes can be specified.

Date to Nominal - The Date to Nominal operator transforms the specified date attribute and writes a new nominal attribute in a user specified format. This conversion is done with respect to the specified date format string that is specified by the date format parameter.

Join - This Operator joins two ExampleSets using one or more Attributes of the input ExampleSets as key attributes. Identical values of the key attributes indicate matching Examples.

Replace Missing Values - This Operator replaces missing values in Examples of selected Attributes by a specified replacement. Missing values can be replaced by the minimum, maximum or average value of that Attribute as well as specific values, i.e. zero.

Forecast Validation - This operator performs a validation of a forecast model, which predicts the future values of a time series.

ARIMA - This operator trains an ARIMA model for a selected time series attribute. The ARIMA operator fits an ARIMA model with given p,d,q to a time series by finding the p+q coefficients which maximize the conditional loglikelihood of the model describing the time series.

Performance (Regression) - This operator is used for statistical performance evaluation of regression tasks and delivers a list of performance criteria values of the regression task. Regression is a technique used for numerical prediction and it is a statistical measure that attempts to determine the strength of the relationship between one dependent variable (i.e. the label attribute) and a series of other changing variables known as independent variables (regular attributes).

Execute Process - This operator can be used to embed a complete process definition of a saved process into the current process definition. The saved process will be loaded and executed when the current process reaches this operator.

Equalize Time Stamps - This operators computes an equalized time series of an input time series with date time indices. The output time series will have new equidistant index values.

Store - This operator stores an IO Object at a location in the data repository. The location of the object to be stored is specified through the repository entry parameter. The stored object can be used by other processes by using the Retrieve operator.

All available operators, as well as their uses can be found in the official rapidminer documentation. (RapidMiner, n.d.)

Even though RapidMiner studio is very powerful it still has some limitations. One of them is number of tweets that can be retrieved within one month. The operator used to retrieve Twitter data has live connection to Twitter API which allows to retrieve only certain number of most recent tweets. This makes it impossible to get few years of historical data which was essential for my research. However, this can be done with Python.

Python is a high-level programming language developed in late 1980s as successor of ABC programming language. Currently, it is one of the most popular programming languages,

used for various kinds of project. Python code can be written in various code editors both online and downloadable and many of them are free to use. For my thesis I decided to use Jupyter Notebook. Jupyter Notebook is a non-profit, open-source project which supports interactive data science and scientific computing across all programming languages. (Jupyter, n.d.)

While Python has large number of libraries for all kinds of projects. A library is a collection of prewritten codes that can be used repeatedly in different programs. As I use Python mainly for retrieving and analysing data, therefore I only use few of those libraries:

Snscrape - snscrape is a scraper for social networking services (SNS). It scrapes things like user profiles, hashtags, or searches and returns the discovered items, e.g. the relevant posts. (JustAnotherArchivist, n.d.)

Pandas - pandas aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. I has a fast and efficient DataFrame object for data manipulation with integrated indexing; tools for reading and writing data, as well as many other data manipulation and analysis tools. (Pandas, n.d.)

Datetime - the datetime module supplies classes for manipulating dates and times. While date and time arithmetic is supported, the focus of the implementation is on efficient attribute extraction for output formatting and manipulation. (Python, n.d.)

Re - this module provides regular expression matching operations. Most regular expression operations are available as module-level functions and methods. (Python, n.d.)

Nltk - NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. (NLTK, n.d.)

Functools - the functools module is for higher-order functions: functions that act on or return other functions. In general, any callable object can be treated as a function for the purposes of this module. (Python, n.d.)

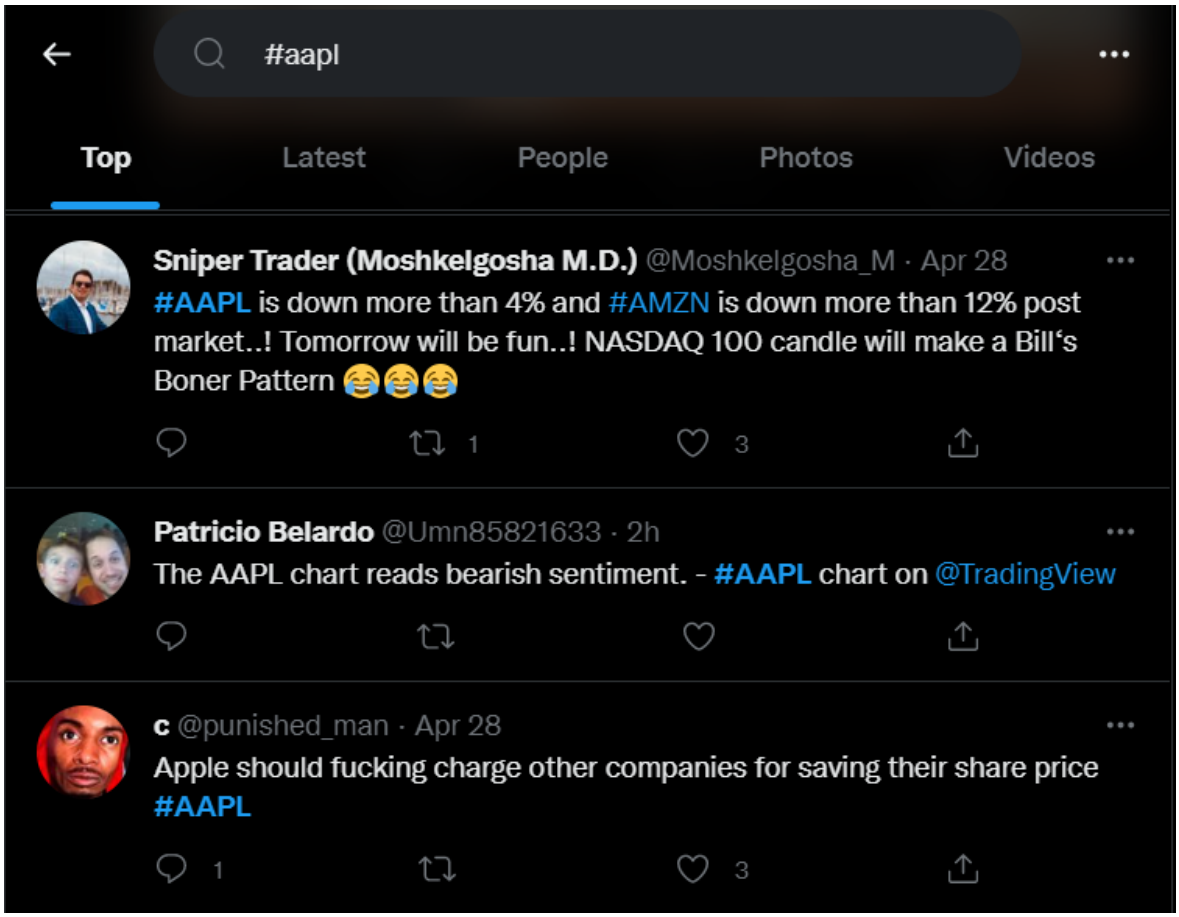
Python (n.d.) provides more information about Python standard libraries are available at. In addition to these, Python has a growing collection of few thousands additional packages.

2.3 Twitter

Twitter is a microblogging and social networking service on which users post their messages and opinions, and those post are referred to as "tweets". Users interact with Twitter through browser or mobile app, or programmatically via its APIs. Number of characters that can be used in one tweet is limited and originally, the limit was set to 140 characters, and now it has doubled to 280. By the start of 2019, Twitter had more than 330 million monthly active users. In reality, the vast majority of tweets are written by a minority of users.

Tweets are publicly visible by default, but senders can restrict message delivery to only their followers. A tweet consists of content of a tweet, hashtags, number of likes and retweets and comments. Hashtags are usually words that best describe the topic of a tweet and they start with special character "#".

Figure 4: Example of Twitter feed

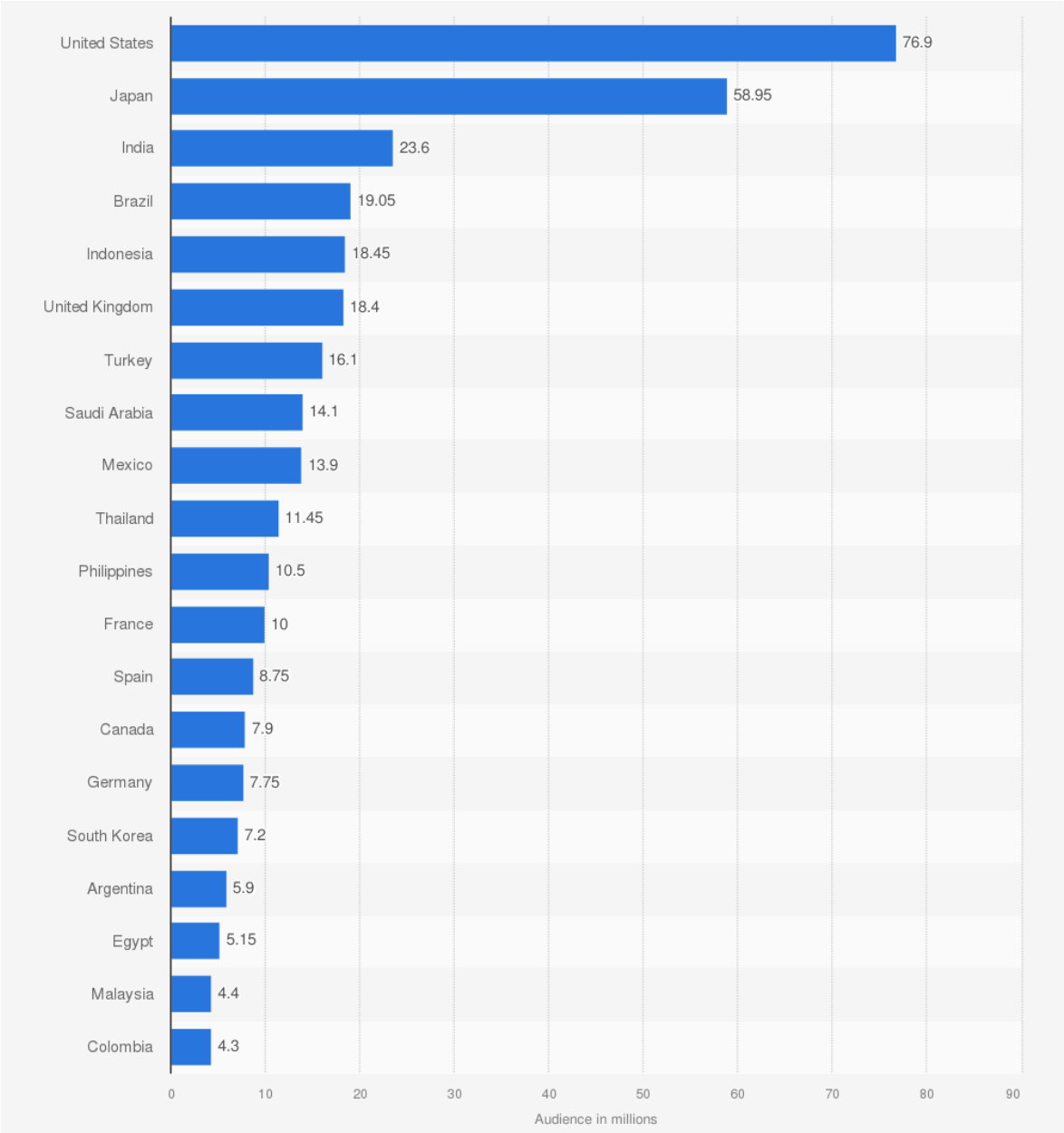


Source: Twitter (2020).

It is important to emphasize that Twitter is not used at the same scale all over the world. Majority of twitter users are from the US, which is a very important fact to have in mind when analysing stock market. This fact influences my decision on the exact market I want to analyse. If I would implement the same sentiment analysis approach on Twitter data on

a company that is more relevant to i.e. stakeholder from South Korea, the results might not affect the price movement as only 7% of South Koreans are actively using twitter to share their opinion, while in the US that share is 77%. Leading countries by number of Twitter users, as of January 2022, are presented in Figure 5.

Figure 5:Leading countries based on number of Twitter users (in millions)



Source: Statista (n.d.).

2.4 Selected market

For my master thesis, I decided to observe the behaviour of stocks of Apple Inc. Apple (ticker symbol: AAPL) is one of the world's leading consumer electronics and personal computer companies. The company, based in California, USA, was established in 1977 as Apple Computer Inc. In early 2007 it dropped the "Computer" from its name. In its beginnings in the late 1970's, company was only selling hand-made personal computer kits. The company continued to focus on personal computers for the following decades, but in recent years the focus has shifted more to consumer electronics such as the iPhone, iPad and iPod. They now also sell a range of non-electronics products like services and applications, with some of the most prominent being the iCloud, iOS, Mac OS and Apple TV. In addition, the company sells and delivers digital applications and software through its iTunes Store, App Store, iBookstore and Mac App Store. Apple has remained focused on developing its own hardware, software, operating systems and services to provide its customers with the best user experience possible. A significant fraction of the company's efforts also go toward marketing and advertising as it believes such efforts are essential to the development and sale of its products. The company has retail stores around the world, with more than 300 locations as of 2012. Apple has five reportable operating segments: Americas, Europe, Japan, Asia-Pacific and Retail. Despite Apple's market-leading position, the company still faces a number of risk factors, which include changing global economic conditions, fluctuating consumer demand, worldwide-competition and potential supply chain disruptions. (Nasdaq, n.d.)

The main reason why I decided to focus my research on exactly this stock is the fact that it is USA-based company. As shown in Figure 6 from previous subchapter, USA has the highest share of population using twitter and therefore there is a huge amount of content available for analysis.

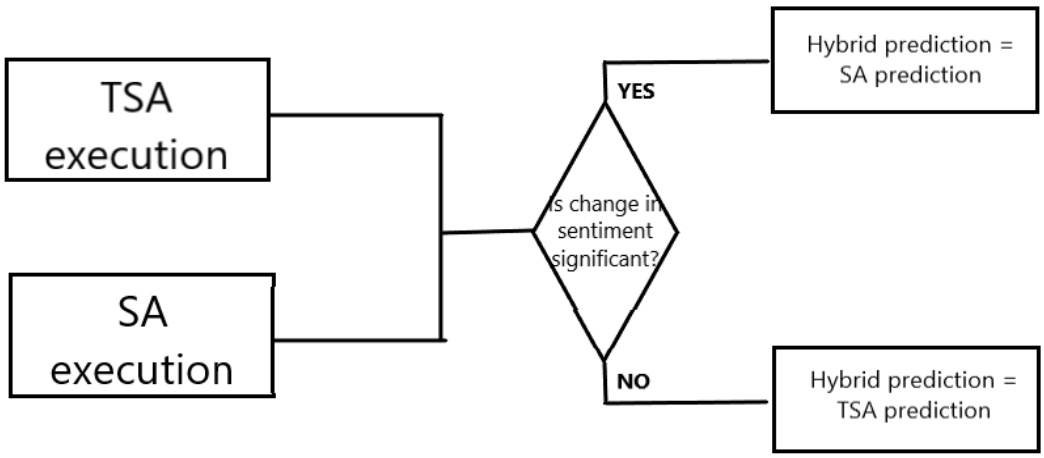
2.5 Hybrid model prototype

"Hybrid models integrate different individual prediction models, which leads to overcoming several limitations of the prediction models in cost of higher complexity of final solution. The main goal is to choose and properly combine a set of prediction techniques in a way of improving accuracy of the final prediction. The improvement of accuracy is achieved by combining advantages of individual prediction models and minimizing their disadvantage." (Rozinajova et al, 2018)

As it is explained in the definition of hybrid model, any combination of two or more models is called hybrid. In the case of stock market prediction, it doesn't necessarily have to mean combination of different types of models. It is also possible to combine, i.e. two different TSA models. No matter which models are combined, the main purpose of their combination is to create a new model which will in the end give better results compare to the individual uses of those combined models.

In the chapters 1.4.1 and 1.4.2 I explained the principles of exact TSA and SA approaches that I use in my thesis. However, hybrid model is more wide and can be implemented differently by each individual. In my thesis it is considered to be a combination of TS ARIMA model and Lexicon-based model for SA. As TSA models are very good in capturing the seasonality and market trends, while SA models on the other hand capture sudden market changes, hybrid model uses best of both of them. This way it can easily be used for any market and in any circumstances. Figure 6 shows the prototype of the hybrid model implemented in this paper.

Figure 6: Hybrid model prototype



Source: Own work.

3 IMPLEMENTATION OF SUGGESTED APPROACHES

This chapter focuses on the whole process of implementing proposed models. As explained in CRISP-DM methodology, in order to implement data mining process, one first needs to understand and prepare the data. So, in the first subsection of this chapter, I explain how did I collect the data, what does the data I collected cover, and how did I clean it. Following three subsections are concentrated on the modelling of SA, TSA and hybrid approach.

3.1 Data collection and preparation

According to CRISP-DM Methodology, after defining and understanding business problem, next steps of data mining process are understanding and preparing the data. This subchapter is concentrated around these steps.

In order to implement suggested models for prediction of stock prices movement, I need both quantitative and qualitative data. The first one are prices of selected stocks which are used in TSA, and second one are tweets about those stocks which are used for SA.

3.1.1 Historical prices

In order to retrieve historical prices of the stocks, I have used Yahoo Finance database. As I have chosen to analyse AAPL, its prices are provided by Yahoo Finance (n.d.). Yahoo Finance has widely available data which is well structured and consists of the most relevant information: open, high, low, close, adjusted close price and volume. Data layout is presented in Table 1. In this research I choose the closing price to represent the price of the index to be predicted, as it reflects all the activities of the index in a trading day. The data used is for period from January 2019 to December 2020. The reason behind the decision to choose exactly this period is the fact that it includes both normal and unexpected circumstances. Here I refer to pre-COVID-19-pandemic and COVID-19-pandemic period. As mentioned before and it is also notable in Table 1, Yahoo Finance provides various daily information about stock price. The high and low columns refer to the maximum and minimum prices in a day. Open and close are the prices at which a stock began and ended trading in the same period. The adjusted closing price is similar to stock's closing price, but it takes into account any corporate actions like stock splits, dividends and rights offering. Volume is the total amount of trading activity. For the purpose of implementing TSA, I will only use information about Closing price.

Table 1: Example of historical prices data

Date	Open	High	Low	Close	Adj Close	Volume
02.01.2019	38.722500	39.712502	38.557499	39.480000	38.277523	148158800
03.01.2019	35.994999	36.430000	35.500000	35.547501	34.464802	365248800
04.01.2019	36.132500	37.137501	35.950001	37.064999	35.936081	234428400
07.01.2019	37.174999	37.207500	36.474998	36.982498	35.856091	219111200
08.01.2019	37.389999	37.955002	37.130001	37.687500	36.539616	164101200
09.01.2019	37.822498	38.632500	37.407501	38.327499	37.160126	180396400
10.01.2019	38.125000	38.492500	37.715000	38.450001	37.278904	143122800
11.01.2019	38.220001	38.424999	37.877499	38.072498	36.912888	108092800
14.01.2019	37.712502	37.817501	37.305000	37.500000	36.357834	129756800
15.01.2019	37.567.501	38.347.500	37.512.501	38.267.502	37.101.959	114843600
16.01.2019	38.270.000	38.970.001	38.250.000	38.735.001	37.555.218	122278800
17.01.2019	38.549.999	39.415.001	38.314.999	38.965.000	37.778.206	119284800
18.01.2019	39.375.000	39.470.001	38.994.999	39.205.002	38.010.899	135004000
22.01.2019	39.102.501	39.182.499	38.154.999	38.325.001	37.157.703	121576000
23.01.2019	38.537.498	38.785.000	37.924.999	38.480.000	37.307.983	92522400
24.01.2019	38.527.500	38.619.999	37.935.001	38.174.999	37.012.264	101766000
25.01.2019	38.869.999	39.532.501	38.580.002	39.439.999	38.238.739	134142000
28.01.2019	38.947.498	39.082.500	38.415.001	39.075.001	37.884.865	104768400
29.01.2019	39.062.500	39.532.501	38.527.500	38.669.998	37.492.191	166348800
30.01.2019	40.812.500	41.537.498	40.057.499	41.312.500	40.054.203	244439200

Source: Yahoo Finance (n.d.).

3.1.2 Twitter data

For SA, I collected twitter data. Twitter is a social media network where users post their opinions about certain topics - tweets. In order to obtain the data I implemented python script in Jupyter Notebook shown in Figure 7. It retrieves all tweets between two dates which contain a specific word, or in my case specific tag. Here, the period of data is shorter as the whole process is highly time-consuming. This means that the script for longer period would be running very long and would probably throw runtime exception. As TS model, which is based on data from yahoo, will need to be divided in training and testing parts it makes sense to use longer period. On the other hand, I will later be finding correlation between sentiment of tweets with prices and I will not be splitting the data, so shorter period should not be problematic.

Figure 7: Scraping tweets with python

```
import snscrape.modules.twitter as sntwitter
import pandas as pd
```

```
# Creating list to append tweet data to
tweets_list = []
for i,tweet in enumerate(sntwitter.TwitterSearchScraper('$AAPL since:2019-12-01 until:2020-06-01 lang:en').get_items()):
    if tweet.user.followersCount >= 10000:
        tweets_list.append([tweet.date, tweet.id, tweet.content, tweet.user.username])
```

```
tweets_df = pd.DataFrame(tweets_list, columns=['Datetime', 'ID', 'Text', 'Username'])
```

```
tweets_df['Datetime'] = tweets_df['Datetime'].dt.date
```

```
tweets_df.to_excel('tweetsAAPL.xlsx',index=False)
```

Source: Own work.

First part of code presented in Figure 7 imports python libraries which contain functions that are used later in the process of retrieving and storing data. Those would be: snscrape and pandas, to be exact twitter module from snscrape. Both are software libraries written for Python as introduced previously.

Then, using twitter search scraper I define variables based on which I want to scrape tweets. In my case those would be search term - $AAPL, starting date – 01.12.2019., ending date – 01.06.2020 and language – EN (English). A suggested by Abbes (2016) I decided for my keyword to start with a $ symbol followed by stock code, as it is very common way of referring to stock in Twitter posts. This convention is very similar to the hashtags and is helpful in order to be sure that the tweet is referring to financial subjects. Additionally I decided to scrape only tweets written in English, as the lexicon I later use for SA has only words in English. Additional filter not applied directly in the scraper, is a lower limit on number of followers of users posting tweets. In this way, I limit the tweets that have higher possible reach, and consequently would have higher impact on overall sentiment. For all of tweets that are collected, I retrieve their posted date, id, content and username of a user who posted that tweet. After I have retrieved all the tweets, I convert the list into pandas dataframe – two-dimensional tabular data structure, and make sure that Datetime attribute is written in the correct format. I save it all in one excel file which I later use for analysing sentiment of those tweets. The layout of twitter data is shown in Table 2.

As mentioned in Methodology and tools used part, this same process could have been done using RapidMiner studio with twitter API, however number of tweets which can be retrieved is very low. As I needed very large volume of data, this workaround was the one allowing me to scrape high amount of tweets in relatively short period of time.

Table 2: Retrieved tweets layout

Datetime	ID	Text	Username
2020-01-03	1213004271102420000	Samsung Shipped 6.7M Smartphones With 5G In 2019, Exceeding Its Own Expectations $AAPL https://t.co/lXOJmdG8u2	newsfilterio
2020-01-03	1212981323775690000	Did $AAPL 300 really mark a short term top for the market? Personally I don't think so. The swing target stays at 3300 which is the first reasonable turning point on big timeframes	SPXTrades
2020-01-03	1212975201748730000	'All I Want To Do Is Run My Own Little PT Boat:' HBO Chief Executive Strikes A Deal With Apple .. $AAPL $T $AMZN $NFLX $DIS https://t.co/1QH0LaeMAW	newsfilterio
2020-01-03	1212974812995430000	Stock Market Today: Sell Ford, Buy Tesla?; More Records $AAPL $TSLA $F https://t.co/uY3Eotmcz9	TopStockAlerts1
2020-01-03	1212969235778900000	Short $AAPL https://t.co/eyXzx2iEBG	MalibuInvest
2020-01-03	1212968247966000000	The following are the top stories in the Wall Street Journal. Reuters has not verified these stories and does not vouch for their accuracy.. $AAPL https://t.co/qjBsnbckxM	newsfilterio
2020-01-03	1212960377408880000	I think the internet is broken because it keeps telling me /ES futures are down. How is $AAPL supposed to gain $100B in mkt cap a month if futures are down. 🤦♂	GS_CapSF
2020-01-03	1212957515463430000	@charliebilello I'm going to assume $AAPL is the best performing stock over the past 40 years. I wonder who is second?	JonahLupton
2020-01-03	1212956670592270000	U.S. stock-index futures fell after a U.S. airstrike near Baghdad international airport killed a top Iranian commander. $AAPL https://t.co/EIsz5UEkn0	newsfilterio
2020-01-03	1212955872445550000	U.S. Stock Futures Fall After Airstrike Kills Iranian Commander $AAPL https://t.co/EIsz5UEkn0	newsfilterio

Source: Twitter (2020).

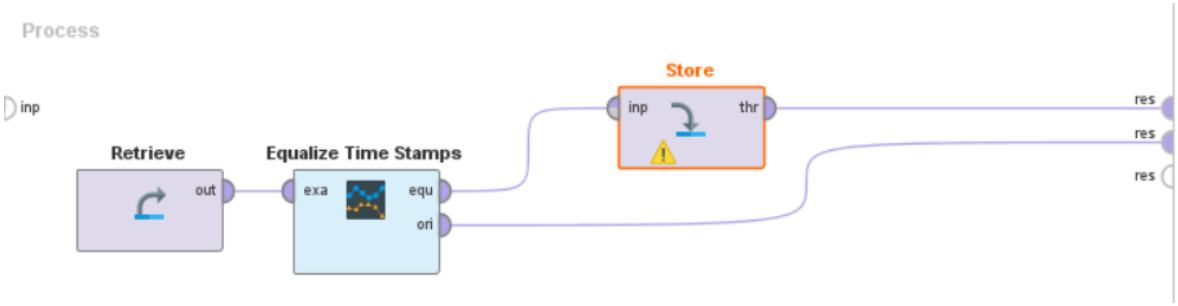
3.2 Modelling

Fourth step in CRISP-DM Methodology is modelling. Here I explain the technical aspects of implemented models. As I have three different models, in following three subchapters, I will in detail describe how I implemented those models.

3.2.1 Time Series Analysis model

TSA was done by using only RapidMiner Studio. As seen in Table 1, stock market is open only during the working days, Monday to Friday. This can be a problem for ARIMA as the data is not equidistant, so the model will have gaps in forecasted values. In order to avoid this issue, I first implemented simple RapidMiner process which assigns closing prices information to missing days – Saturdays and Sundays. The process is shown in Figure 8. It retrieves data downloaded from Yahoo Finance and for every missing day it assigns the price of last available day. This means that Saturdays and Sundays will have closing price of Friday, as price will not change over the weekend.

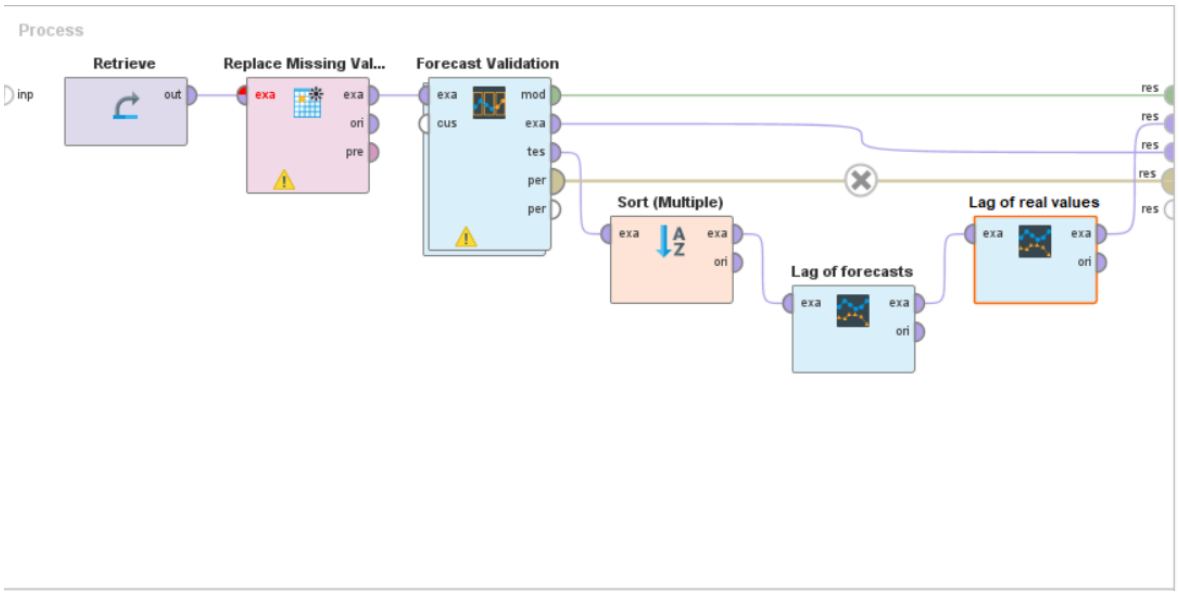
Figure 8: Equidistant historical stock prices data



Source: Own work.

This is a very important step, especially when using ARIMA to implement TSA. Now that I have equidistant data, I can continue with implementation of ARIMA process as shown in Figure 9.

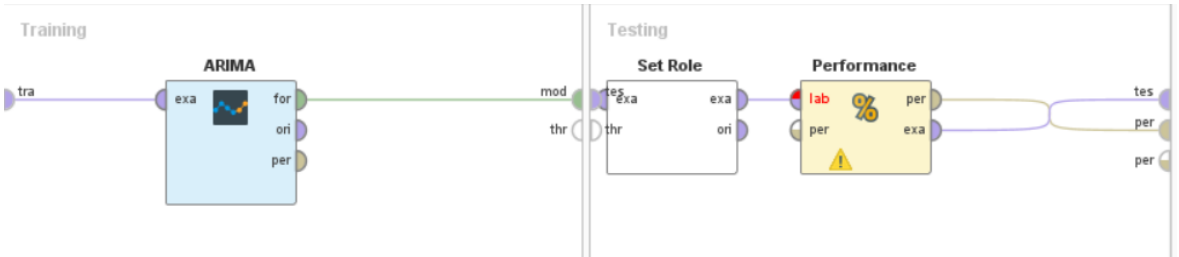
Figure 9: Forecasting with ARIMA



Source: Own work.

First step of the process is to retrieve the equidistant data on prices of a certain stock. As the data is already clean, I only need to decide how to deal with missing values. What I did was to replace missing values by average closing price. Even though Yahoo data is of high quality and does not have missing values, except weekend values as explained before, this step still can't be skipped in order to have high-quality process even in case the source data changes. Then, after the data is completely clean, I proceed to forecasting part by using Forecast validation window. Here, I set the attribute Close, which represents closing value of stock for each day, to be the time series attribute. This means that this attribute will be forecasted. I also set the window size to be at 80% of all available data. With this, I divide the data in a way that 80% will be used for training, and 20% for testing. The forecasting process itself is inside this window and is shown in Figure 10.

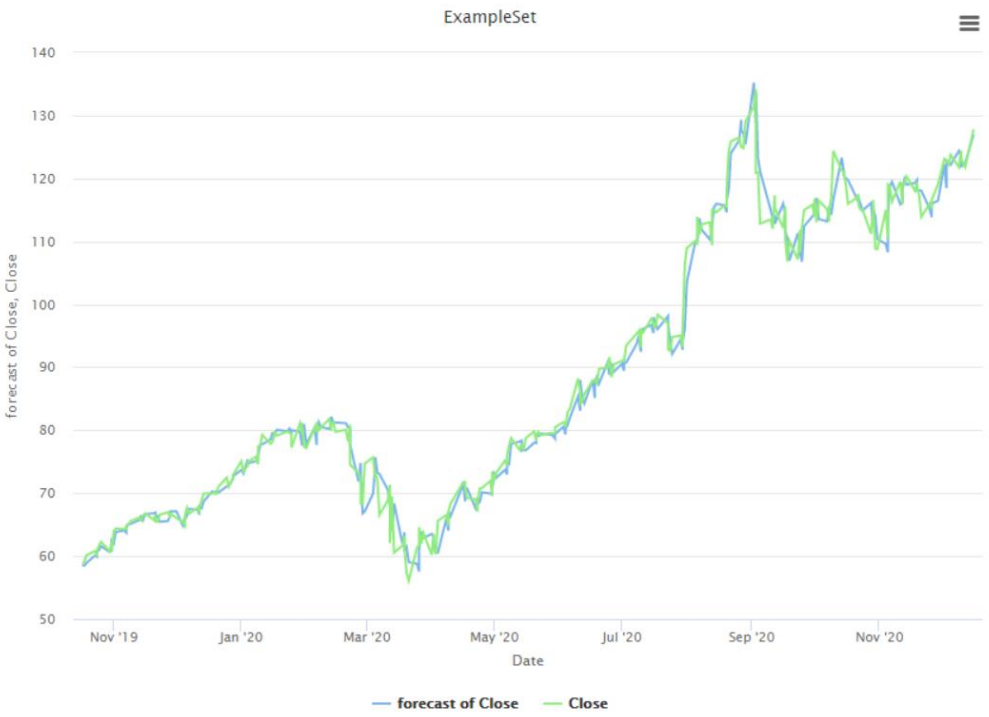
Figure 10: Forecast validation window



Source: Own work.

ARIMA operator trains ARIMA model for a Close attribute and uses Date as indices attribute. After testing different p, d, q values, (5, 0, 2) combination yield with best performance with relative error of 1.89% +/- 2.02%.

Figure 11: Difference of real and forecasted prices for AAPL



Source: Own work.

Further, I sort the dataset by date and for each row I save the forecasted and real prices of the following days by using Lag operator. I do this, because I will use those values later to calculate whether stock price will rise or fall in the coming days. The process shown in Figure 9 has four outputs: ARIMA model perimeters definition, Performance values, original dataset on which model was executed on and final dataset with forecasted and lagged values.

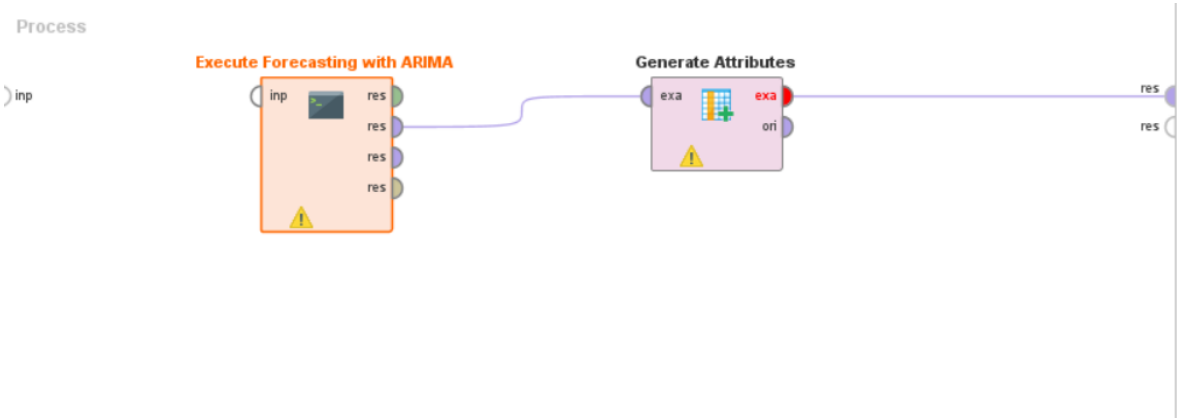
Next part of TSA implementation is designed to forecast the movement of the market. To do this, I use the final dataset with forecasted and lagged values form previous process. Figure 12 shows the next process. What it does is to calculate whether stock price will rise or fall.

The base for this process is ARIMA model explained previously. To build upon it, I generate two new attributes on the final dataset: rise/fall forecast and rise/fall real. The first one calculates whether the price will rise or fall by comparing forecasted values for next 5 days with Close value of that day, while the second one does the same with real Close values for next 5 days. In order to calculate that, I used following formula:

```
if(avg([forecast of Close+1],[forecast of Close+2],[forecast of
Close+3],[forecast of Close+4],[forecast of Close+5])-
Close<Close*0.05, if(avg([forecast of Close+1],[forecast of
```

```
Close+2],[forecast of Close+3],[forecast of Close+4],[forecast of
Close+5])-Close>- Close*0.05,0,-1),1)
```

Figure 12: Rise/fall process



Source: Own work.

This means that if the difference between average of forecasted prices for following 5 days and the closing price of that day is less than 5% of current Closing price and is greater than -5% of current Closing price, rise/fall forecast attribute will be assigned value 0 – meaning that there should be no significant change in price (price will stay approximately the same). If the difference is greater than 5% of current Closing price, this attribute will be assigned value 1, meaning the price will rise, and otherwise it will have value -1 and conclusion is that the price will fall. It is important to emphasize here that the threshold value, 5% of current Closing price, can be adjusted for different stocks.

The other attribute – rise/fall real is calculated in the same way, using real values instead of forecasted ones. Later, I will use this attribute to calculate the performance of forecasted one.

3.2.2 Sentiment Analysis model

In order to analyse sentiment of retrieved tweets, I have first implemented python script shown in Figure 13 as it doesn't have limitations on number of tweets that could be analysed.

The script below first imports all libraries needed to clean the data and analyse sentiment. Then I define a function "cleanTxt", used for data cleansing. What it does is to transform all letters to be lower case, and then removes username mentions (all words that start with "@"), hashtags (all words that start with "#"), hyperlinks, punctuations and non-alphanumeric characters. Next step in this function is tokenization, meaning that it splits all provided content into list of words. I do this because I want to remove certain words - "for", "on",

"an", "a", "of", "and", "in", "the", "to", "from", which don't have any effect on sentiment of the text. The last step is to put all filtered words together, separated by a blank space.

Next part of the script reads previously extracted tweets from an excel file, applies above defined cleanTxt function on the content of tweets and extracts polarity. Polarity score ranges from -1.0 to 1.0 which gives an indication of weather overall sentiment is positive, negative or neutral in the case of polarity score 0.

In order to extract polarity, I use polarity_scores function from VADER's SentimenIntensityAnalyser. As explained previously, VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based SA tool that is specifically used to analyse sentiments expressed in social media. VADER uses a sentiment lexicon which is a list of lexical features (words) which are generally labelled according to their semantic orientation as either positive or negative. VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is (GeeksForGeeks, n.d.).

The last step is to save the information about average polarity score, total number of tweets, and number of negative and positive tweets for each observed day. I consider a tweet to be negative if it has polarity score <= -0.1 and positive if the score is >= 0.1. Data layout of this process is shown in Table 3.

All this information is stored in an excel file which I later use to find correlation between extracted information and closing stock prices. The RapidMiner process shown in Figure 14 does exactly that.

First step in this process is to retrieve sentiment data which has previously been stored in an excel file. I then calculate the shares of positive and negative tweets by simply dividing number of positive or negative tweets over total number of tweets for the day. The initial idea was to predict whether price of stock will rise or fall based on average daily sentiment. However, average daily sentiment was relatively constant in this period and was leaning towards positive overall sentiment, so there wasn't strong correlation between Average Sentiment and daily closing price. This is why I changed the approach slightly. As I have already explained, not only did I extract the data about average sentiment scores, but as well number of positive, negative and overall daily tweets. Based on these information I can try to find correlation between shares of positive and negative tweets with closing prices. The next step from Figure 14 is to calculate difference of the shares of positive and negative tweets. As stock market is closed on weekends – Saturdays and Sundays, the sentiment data for these days will be lost after joining this dataset with prices. The reason why I don't use equidistant prices data as in TSA is because the changes over weekend are reflected only on Monday. If I would use equidistant data, this model would do the predictions for Saturdays and Sundays as well, but they would not be relevant and could potentially skew Monday predictions. ARIMA model itself requires equidistant data in order to make proper

predictions, but that is not a condition in SA, and would rather be possibly harmful in this process. But this data is still valuable as it can be reflected on Monday prices. This is why for each day, by using Lag operator, I also retrieve share differences of two following days and I add additional column in the dataset which shows the name of each day. Then I change the share difference attribute for Friday to be average of share difference of Friday, Saturday and Sunday.

Figure 13: Sentiment Analysis with Python

```
import pandas as pd
import datetime as dt
import re
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

```
nltk.download('vader_lexicon')
```

```
def cleanTxt(text):
    tempTxt = text.lower()
    tempTxt = re.sub("'", "", tempTxt) # to avoid removing contractions in english
    tempTxt = re.sub("@[A-Za-z0-9_]+","", tempTxt)  # to remove mentions
    tempTxt = re.sub("#[A-Za-z0-9_]+","", tempTxt)  # to remove hashtags
    tempTxt = re.sub(r'http\S+', '', tempTxt)  # to remove links
    tempTxt = re.sub(r'www\S+', '', tempTxt)  # to remove links
    tempTxt = re.sub('[()]', ' ', tempTxt)
    tempTxt = re.sub('\[.*?\]',' ', tempTxt)
    tempTxt = re.sub(r"(.)\1+", r"\1\1", tempTxt) # to remove more than  occurences of the same letter in a word
    tempTxt = tempTxt.split()
    stopwords = ["for", "on", "an", "a", "of", "and", "in", "the", "to", "from"]
    tempTxt = [w for w in tempTxt if not w in stopwords]
    tempTxt = " ".join(word for word in tempTxt)
    return tempTxt
```

```
tweets=pd.read_excel('tweetsAAPL.xlsx')
```

```
tweets['Text'] = tweets['Text'].astype(str)
```

```
tweets['Text']=tweets['Text'].apply(cleanTxt)
```

```
sid = SentimentIntensityAnalyzer()
```

```
tweets['scores'] = tweets['Text'].apply(lambda Text: sid.polarity_scores(Text))
```

```
tweets['SentimentScore']  = tweets['scores'].apply(lambda score_dict: score_dict['compound'])
```

```
tweetsDailyPolarity1 = tweets.groupby('Datetime', as_index=False, sort=False)['SentimentScore'].
mean().rename(columns={'SentimentScore':'AverageSentimentScore'})
```

```
tweetsDailyPolarity2 = tweets.groupby('Datetime', as_index=False, sort=False)['SentimentScore'].
count().rename(columns={'SentimentScore':'NumberOfTweets'})
```

```
tweetsDailyPolarity3 = tweets.groupby('Datetime', as_index=False, sort=False)['SentimentScore'].
apply(lambda x: x[x<=-0.1].count()).rename(columns={'SentimentScore':'NegativeTweets'})
```

```
tweetsDailyPolarity4 = tweets.groupby('Datetime', as_index=False, sort=False)['SentimentScore'].
apply(lambda x: x[x>=0.1].count()).rename(columns={'SentimentScore':'PositiveTweets'})
```

```
from functools import reduce
```

```
tweetsDailyPolarity = reduce(lambda x,y: pd.merge(x,y, on='Datetime', how='outer'),
                             [tweetsDailyPolarity1, tweetsDailyPolarity2, tweetsDailyPolarity3, tweetsDailyPolarity4])
```

```
tweetsDailyPolarity.to_excel('DailyPolarityAAPL.xlsx',index=False)
```

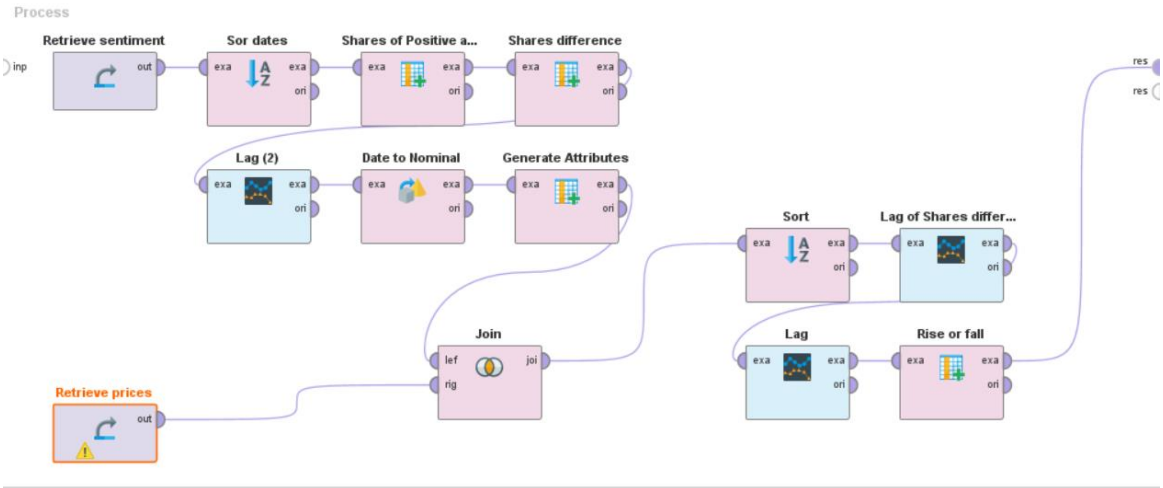
Source: Own work.

Table 3: Data layout of daily sentiment data

Datetime	AveragePolarity	NumberOfTweets	NegativeTweets	PositiveTweets
2020-01-03	0.138291	38	0	4
2020-01-02	0.119688	306	2	31
2020-01-01	0.141687	49	0	4
2019-12-31	0.118899	176	4	21
2019-12-30	0.150923	228	2	29
2019-12-29	0.102478	53	1	4
2019-12-28	0.116289	45	0	5
2019-12-27	0.126840	221	2	24
2019-12-26	0.133369	177	1	18
2019-12-25	0.196605	28	1	8
2019-12-24	0.160107	97	1	12
2019-12-23	0.115596	200	0	12
2019-12-22	0.131113	53	1	5
2019-12-21	0.207845	41	0	6
2019-12-20	0.114581	172	1	20
2019-12-19	0.101953	119	1	9
2019-12-18	0.105229	154	0	13
2019-12-17	0.096665	207	1	10
2019-12-16	0.119721	182	1	13
2019-12-15	0.109908	58	0	6

Source: Own work.

Figure 14: Analysing sentiment scores with RapidMiner



Source: Own work.

After manipulating the data, next step is to join this dataset with the one containing daily price information. The two datasets are joined by date. I then sort them from oldest to newest date. I assume that changes in shares will not directly impact that day's closing price. So,

again, by using Lag operator, I save values of shares differences of three previous days. The last step is to save difference of difference of positive and negative tweets share and average of those differences of three previous days. This parameter shows if there has been strong change in the shares of positive and negative tweets on that day and will help me to make predictions with hybrid model.

By running simple correlation (Pearson product-moment correlation coefficient) I noticed that share difference of third previous day had strongest correlation of 0.29 with closing price. The Pearson product-moment correlation coefficient measures how strong is linear association between two variables, in this case closing price and share difference of positive and negative tweets of third previous day. In principle, it tries to draw a line of best fit through the data of two variables, as shown in Figure 15. The coefficient indicates how far away from the line of best fit are all data point which were used to analyse correlation. It takes values between -1 and 1. Value 0 indicates that variables observed are not correlated, while values greater than 0 suggest positive correlation and values less than 0 suggest negative correlation. Positive correlation means that as the value of one variable increases, so does the value of the other variable. On contrary, negative correlation indicates that if value of one variable increases, the other one decreases and vice versa. (Laerd Statistics, n.d.)

Even though the correlation is not significantly strong, it is clear from Figure 16 that this data does have an impact on the prices. Table 4 shows correlation values between Closing price and different Sentiment-Based attributes.

Table 4: Correlation of closing prices with different sentiment attributes

Attribute	**Correlation with Closing price**
Average Daily Sentiment Score	0.16
Difference of positive and negative tweets of that day	0.23
Difference of positive and negative tweets of one previous day	0.27
Difference of positive and negative tweets of two previous days	0.25
Difference of positive and negative tweets of three previous days	**0.29**

Source: Own work.

Figure 15: Pearson correlation coefficient - line of best fit

Source: Own work.

Figure 16: Impact of tweets on closing prices

Source: Own work.

In the Figure 16, I depict the graph with daily closing prices and difference between shares of positive and negative tweets of three previous days. There are two things which can be observed here. First, even though tweet-related line has more noise, both lines still follow similar pattern. Secondly, it is notable that when the difference of shares sharply falls,

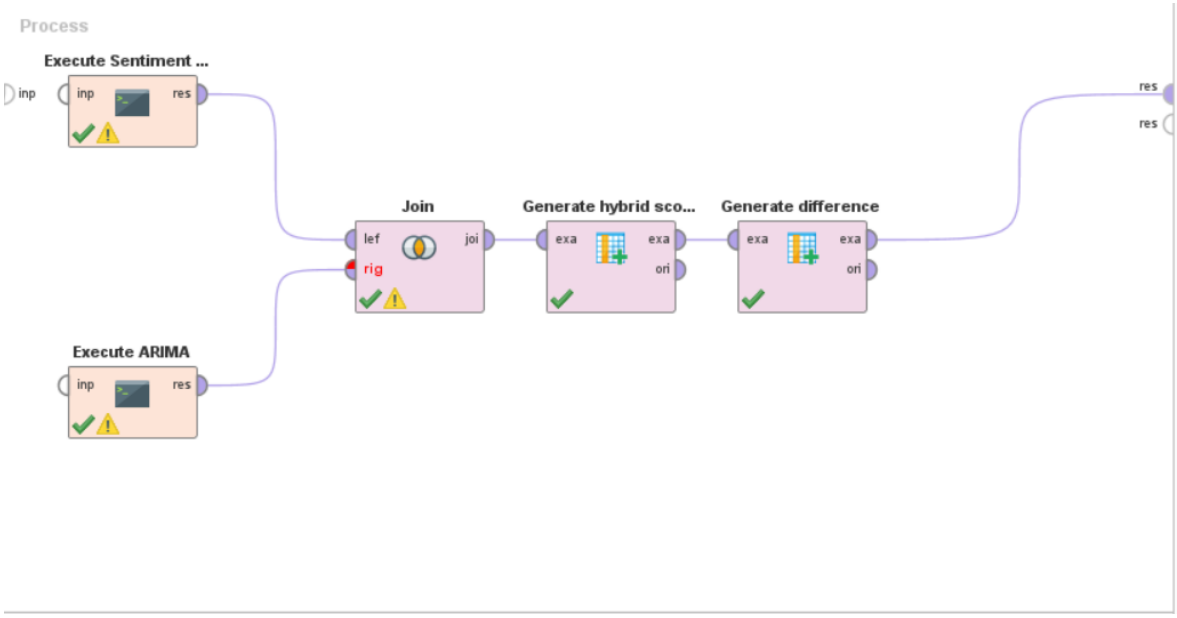
meaning that share of negative tweets increases, closing price of stock starts to go down. Similarly, sudden increase in the difference of shares can be followed by increase in closing price.

3.2.3 Hybrid model

As explained previously, hybrid model in general is a combination of two or more models created that makes use of benefits of all combined models in order to create more efficient one. In my thesis, hybrid approach uses SA to improve TS predictions. As it is shown in chapter 3.2.1 ARIMA-based TS approach can be good for stable markets, but unexpected market changes are drawback of this model. However, it is also shown in chapter 3.2.2 that changes in shares of positive and negative tweets could have an impact on the direction of market prices. That is why it is reasonable to assume that combination of the two could be more efficient for both stable and unstable markets. Similar to what Kesavan, Karthiraman, Ebenezer & Adhithyan (2020) did when analysing Indian stock market, I too will combine information on historical prices with shares of positive and negative tweets. As ARIMA model showed very promising results for AAPL stock, I decided to use this model as a base and on top of it to add SA in the occasions when there is a strong change in sentiment towards market.

The process, implemented in Rapid Miner, can be seen in Figure 17.

Figure 17: Hybrid approach process in RapidMiner



Source: Own work.

First step of this process is to Execute Sentiment Analysis and ARIMA Forecasting processes which I have previously explained in chapters 3.2.1 and 3.2.2. I then join them

based on date attribute. As I already have information about ARIMA predictions, I now have to combine that information with the relevant information from SA.

I previously elaborated on the fact that difference in shares of positive and negative tweets of 3 previous days has the strongest correlation with stock prices. That is why I take that information and check whether there has been significant change in the shares. If there has been increase of more than 30 percentage points, meaning that either the share of positive tweets had risen or the share of negative tweets had fallen, I consider it to have very positive impact. The same way, I believe that decrease of more than 30 percentage points, has very negative impact. Threshold of 30 percentage points can be increased or lowered directly in the process if it is suitable for the observed market. In this case, when observing AAPL stock, after testing different threshold values, the value of 30 has proven to be the most relevant.

After joining the processes, I generate the prediction of new, hybrid model. The idea here is to take the predictions of TS ARIMA model as a base and on top of them add predictions of SA model, only in the occasions when they are relevant – meaning that there is a strong change in overall opinion towards market. So, in the days when there is significant change in public opinion towards a stock, hybrid score is taking it into consideration and has value 1 if the change is towards more positive sentiment and -1 if it is inclining towards more negative sentiment. Otherwise, when public opinion is more stable, and change in sentiment is not significant, hybrid score is same as ARIMA score.

4 EVALUATION

After implementing the models and making sure they work correctly, next step of CRISP-DM methodology is to evaluate them. In the evaluation part, I make sure that previously set goals are reached and that research question, which represented a business problem, is answered. As stated in the Introduction, goals of my thesis are:

- To explore which approach for stock market prediction, SA or TSA, gives better results
- To implement hybrid approach of the two and investigate if the accuracy of its prediction would be higher.

Chapter 4.1 Analysis of results elaborates on these goals. In this subchapter, I present the results of implementation of all three models which give an indication on whether the goals were set. Then in chapter 4.2 Answering the research question, I elaborate on the exact answer to my research question stated in the beginning:

“Is there a significant difference in accuracy of combination of two stock market prediction approaches: market sentiment analysis and time series analysis in comparison with their individual uses?”

4.1 Analysis of results

In this part of my thesis, I observe the performance of all three previously implemented approaches and make the comparison between them. Based on their performance, I elaborate on the set goals. Since I implemented three different approaches, this section is divided into three subsections.

4.1.1 Time Series Analysis results

In the implementation part of TSA, I commented on the performance of implemented ARIMA model which has relative error of 1.89% +/- 2.02%. The average relative error is the average of the absolute deviation of the prediction from the actual value divided by actual value. This means that on average, predicted values were around 2% different than the actual ones. This is a very high prediction accuracy, but we need to have in mind that AAPL is relatively stable market, and as such, it is very suitable for predictions with ARIMA.

However, I also want to check the accuracy of predicting the movement of the market. In ordered to do that I compared two variables implemented in the model: rise/fall ARIMA and rise/fall real. Both of them have three possible values: 1 if difference between average of closing prices of next five days and that day's closing price is higher than 5% of the day's closing price, -1 if that difference is less than -5% and 0 otherwise. One attribute uses forecasted prices of ARIMA and compares them to that day's real closing price, while the other uses real values. Attribute based on ARIMA was equal to the one based on real values in 90% of days. This accuracy is still very high, and one reason for this could be because this market isn't very volatile. But still, this accuracy is lower than accuracy of prices prediction. This is because ARIMA is constantly being adjusted with each iteration, so even though it will make wrong prediction when market suddenly falls, it's prediction for the following days will be adjusted based on that fall. However, when making a decision about whether to invest in market or not, one wants to know in advance if the market will of will not go up/down.

In my research paper, I found the change in price of 5% is relevant for making a decision for investment. However, this 5% change can always be adjusted based on either stock for which market movement is being predicter or one's personal needs and aspirations. It is only about adjusting the formula for rise/fall attributes, introduced in modelling part of TSA, chapter 3.2.1.

4.1.2 Sentiment Analysis results

As explained in the implementation part, for AAPL stock I didn't find high correlation between twitter data and closing prices. Therefore, when I tried to make predictions by only using sentiment parameters, results were very different when setting various thresholds.

Here, threshold is referring to the value defined to be significant change in shares of positive and negative tweets. Similarly, as in the paper of Smailovic, Grcar, Lavrac, & Znidarsic (2013) who were observing daily change in number of positive tweets, I have also used information on shares of both positive and negative tweets. In the implementation part I have defined an attribute for difference of the shares of positive and negative tweets. Then I calculated the change of this attribute compared to its values of three previous days. If I compare this change to different thresholds the results vary, but the accuracy in neither case is very strong. This is why in my opinion, using only SA in the way I implemented it to predict stock price movements is not encouraging. However, there are few more points to discuss.

- First, even though this approach is not to be used for every-day predictions, strong changes in opinion towards certain stock can still define market behaviour. This is clearly depicted in Figure 15
- Second, AAPL is relatively stable market and there are not a lot of big changes in neither prices nor overall sentiment towards it. But, clearly sentiment changes did affect market behaviour and can help in the decision about the investment

Comparing this to TSA approach, it is very hard to say that one approach is better than other because, results of SA are not quantifiable as they are not used in everyday predictions, so their accuracy can't be measured, while TSA results are quantified. Additionally, for stable market as AAPL is, ARIMA makes very precise price predictions, but is not as accurate in predicting unexpected market movements. On the other hands, SA does not predict any changes in every-day situations, but when it comes to sudden price fluctuations it is a key factor to take into account.

4.1.3 Hybrid Approach results

The results of Hybrid approach were observed in similar way to TS results. Even though time series by itself was trained and tested on larger scale of data, in order to make direct comparison of two models, I consider only timeframe in which all approaches are implemented. Table 5 present accuracy of different models.

Table 5: Prediction accuracy

Stock	TS Accuracy	Hybrid Accuracy
AAPL	90%	92%

Source: Own work.

The table shows only accuracy of TS and hybrid model, because as explained earlier, SA is not to be used on daily basis. Looking at the table 5, accuracy of Hybrid approach is higher than TS, however the difference is not significant. This can be result of two things:

- First, ARIMA has very good performance in predicting future values for AAPL stock in the observed period
- Second, price of AAPL didn't have many outstanding changes in this period

These two observations are interconnected. As I have previously explained, ARIMA is making forecasts based on previous values, so it is reasonable that it is highly accurate when making predictions for stable markets, as AAPL is.

To be better assured about the importance of adding SA to ARIMA forecasts, let's focus on one special event. On 18. February 2020, the share of negative tweets was significantly higher than on the previous day causing the difference in shares of positive and negative tweets to go down by 37 points. As I have previously explained, the highest correlation with prices was found for share changes of three previous days. That is why this change is expected to be reflected three days later. From Figure 18 it is clearly visible that the closing price of AAPL stock started going down around 21. February 2020.

Figure 18: Importance of Sentiment Analysis

Source: Own work.

It is clear from the Figure 18 that prices of the stock were very stable in the days before 21. February 2020. This is why ARIMA forecasted that price will continue to follow the same pattern, meaning that there will not be any significant change in the following 5 days. This

further means that rise/fall attribute of TSA which predicts weather price will rise or fall, was assigned value 0. However, the prediction is wrong as price suddenly started going down. But, as the change in difference of positive and negative tweets of 3 previous days was higher than 30 points in favour of negative tweets, hybrid approach here gives SA more advantage. As a result, rise/fall attribute of hybrid approach was assigned value -1, meaning that it expects price to go down in the following 5 days. In this particular case we see that Hybrid approach really did improve TS prediction.

This leads to a similar conclusion as the one from Mehta, Malhar, and Shankarmani, (2021). Considering the fact that ARIMA is linear model, SA can help predicting sudden increases or decreases in the market prices.

Both goals set in the beginning are satisfied, as I did successfully implement all three approaches and concluded that hybrid one gives better picture on market movement than bot TSA and SA used on its own.

4.2 Answering the research question

Finally, after presenting the implementation of different methods for market movement prediction and analysing the results of each of them, it is time to answer the research question presented in the beginning:

"Is there a significant difference in accuracy of combination of two stock market prediction approaches: market sentiment analysis and time series analysis in comparison with their individual uses?"

My research paper was based on the case study of specific company stock – AAPL. Therefore, the conclusion can't be taken as general, but is still gives some general indications on the efficiency of all three prediction models presented. Even though results for a specific stock – AAPL didn't show very significant improvement in the hybrid approach compared to other two approaches, I would still say that this approach is by far the best. Let's observe it from two perspectives: volatile and stable market.

1. Volatile markets are very hard to be predicted by only using TSA, especially linear model as ARIMA is, as they don't have very clear pattern of behaviour. These markets are usually very dependent on external factors which can be reflected in public opinion. As twitter is a social network used by wide range of users, from which many of them are highly influential and can impact the form of overall global opinion towards something, it is very exhaustive source of information. As such, it can reflect the impact of external factors and can give better insight in market behaviour. In this case, adding SA on top of TSA would definitely be more accurate in predicting rise or fall of certain stocks price.

2. For stable markets, TSA can be very accurate as those markets don't have many ups and downs. But, what if there is a sudden increase or decrease in price caused by external factors? TSA is not able to predict this as it is always based on historical patterns. Here SA plays major role as it does reflect those factors. Still using only SA is not a good idea, as market can have seasonal behaviour, or can follow certain trends, and TSA can capture it, while SA can't. As a result combination of the two then includes both seasonality and trend pattern as well as effect of external factors and is in fact the best model to be used for predictions.

So, to answer the research question, for this specific market, the difference in accuracy of hybrid model and TSA is not significantly high, but still hybrid model did improve the accuracy of TSA to some extent. Based on the reasons stated above, it is safely to conclude that in order to make data-based decision on investment, using hybrid model for decision-making process is the safest one and can yield higher returns.

CONCLUSION

My thesis is divided into two main parts: theoretical introduction of the different approaches for stock market predictions and hands-on implementation of three approaches: TSA, SA and hybrid of the two. Even though there were some limitations to the implementation part of my thesis, I managed to reach the goals set in the beginning and to answer my research question.

Stock market prediction has been a topic of interest for many researches for a long time now, and it is likely to be interesting in the future as well. Many models have been implemented in order to predict the future prices of stocks and because of so-called chaos in the market none of the models had, and very likely will never have, 100% accuracy. But researchers are in constant competition to make more and more accurate models. However one of the problems, which is also limitation of my thesis, is that it is very hard to make a general model or to say that one model that fits all and is in general better than all the others. This is because different markets have different behaviour, and the same model can have largely different accuracy for two different stocks. What I have observed is that large proportion of these papers focus on predicting the exact price of the stock. I believe that, especially in the cases when SA is involved, predicting the movement of the market, i.e. whether it will rise or fall over/below certain threshold, is more relevant than knowing what the exact price will be.

In my thesis, I do exactly this. I predict weather the price in following few days will be higher or lower than current price by more than 5%. I first implement two classical approaches, TSA and SA.

For TSA I use ARIMA as it has proven to be highly accurate in prediction of stock prices. I observe it in the case of AAPL which didn't have many price fluctuations in the observed period. In this case, implemented model had very promising with low relative error.

Consequentially prediction of market movement is very high. The final result is that in 90% of observed days, ARIMA-based TSA model predicts correctly if the price of AAPL stock will rise or fall.

For SA, I use lexicon-based approach to evaluate sentiment of each tweet. I combine these sentiments to calculate average daily sentiment for AAPL stock. By using difference in shares of positive and negative tweets I predict if the price of stock will rise or fall. In cases of high daily fluctuations of positive and negative tweets, i.e. if the share of negative tweets rises sharply from one day to other, the model predicts the market behaviour very accurately. However, if there are no significant fluctuations model would predict that the price will not be changed, even though in some cases it is changed. This is why this model is not to be used alone.

Finally, even though TSA shows very good results, they can still be improved, and I do that by adding SA on top of TSA to create a hybrid model of the two. Hybrid model relies on TSA in the times of no big changes in overall sentiment. But if there is a big change in overall sentiment, hybrid model tweaks the prediction of TSA based on SA. The final result really proves that combining these two approaches can predict market movement better than using them on their own. Hybrid model predicted market movement accurately in 92% of observed days.

The implementation of these three models was not completely smooth as I had to deal with some problems. The biggest one is related to SA and is the fact that there is no public database with tweets on different stocks, so I had to do it manually. Initial option was to use Twitter API which is very limited and would not allow me to get all tweets for the whole period I observed. This is why I had to implement new script that can get the data for the whole period. However, the script is very slow and takes a lot of time to execute even if I am having only one search word.

Although I did answer my research question, I still had some limitations while implementing all three models:

1. The VADER lexicon was the best fit for the purpose of mining the sentiment from Twitter. However, as it is final lexicon, meaning that is has limited number of words available for analysis, not all stock-related terms are available in it. This might be one of the reasons why average daily sentiment score was not relevant to be used as parameter for prediction.
2. Twitter feed is mostly written in English language, which can be limitation for some markets. Additionally, Twitter is not as largely used in all countries, so sentiment towards some markets might not be expressed on Twitter.
3. The way I implemented all three models is only usable for short-term predictions. Even though it is good for fast-returns which are usually lower-scale, these models can’t be used for making decision on long-term high-return investments.

Even though there are many research papers in the field of stock market prediction, majority of them focus on predicting the exact future price of a stock. This paper focuses rather on predicting whether price will go over or under the certain threshold, and this relatively insufficiently researched approach is one of the added values of my thesis. Also, while the focus of many SA-based papers is on the whole twitter feed, this one limits the user based on number of followers to avoid the tweets that would have low reach and therefore are hardly to impact price movement. Both SA and TSA, and therefore hybrid model, can be adjusted to different markets, as thresholds that are set and which are basis for prediction result are easily changed based on the market observed. Lastly, implemented approaches can be adjusted to reflect return on investment aspired.

In the future work, I would suggest including more keywords while scraping twitter as it can give broader sentiment. Additionally, I believe it could be beneficial to upgrade VADER lexicon with supplementary lexicon which would include sentiments of stock market related words. Furthermore, it would be good to test the models with more thresholds on what is considered to be strong change in overall sentiment. Testing this could improve the accuracy of SA and consequentially hybrid model as well. Also, it would be good to test the results on different kinds of markets, i.e. volatile and stable as well as on markets from different regions. This could indicate if there are some regional markets that are more reliant on SA in which cases hybrid approach could be very beneficial for making short-term investments. Lastly, I would suggest exploring the correlation of SA over long period of time with long-term forecasts.

REFERENCE LIST

1. Abbes, H. (2016). Tweets sentiment and their impact on stock market movements. https://matheo.uliege.be/bitstream/2268.2/1323/5/Thesis_HakimAbbes_s110997.pdf
2. Adebiyi, A. A., Adewumi, A.O. & Ayo, C.K. (2014). Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. *Journal of Applied Mathematics*, 2014(1),1-7.
3. Alestair, V., Harpreet, T., Aquib, S., Prasenjit, B. & Ashish, R. (2016). Stock market prediction using Time Series. *International Journal on Recent and Innovation Trends in Computing and Communication*, (4), 427-430.
4. Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Stock Price Prediction Using the ARIMA Model. *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation,* 106-112.
5. Azhikannickel, J. (2019). *Time Series Analysis on Stock Market Forecasting (ARIMA & Prophet)*. Medium. Retrieved 20 December 2021 from: https://medium.com/@josephabraham9996/time-series-analysis-on-stock-market-forecasting-arima-prophet-2b60cacf604
6. Bharathi.Sv, S. & Geetha, A. (2017). Sentiment Analysis for Effective Stock Market Prediction. *International Journal of Inelligent Engineering and Systems*, 10(3), 146-154.

7. Chauhan, N.S. (2020). *Stock Market Forecasting Using Time Series Analysis*. KDnuggets. Retrieved 16 September 2021 from: https://www.kdnuggets.com/2020/01/stock-market-forecasting-time-series-analysis.html
8. Chou, H-C. (2021). *Combining Time Series and Sentiment Analysis for Stock Market Forecasting*. Graduate Theses and Dissertations. Retrieved 25 February 2022 from: https://digitalcommons.usf.edu/etd/8749
9. Corporate Finance Institute. (n.d.). *Efficient Markets Hypothesis*. Retrieved 5 May 2022 from: https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/efficient-markets-hypothesis/
10. Data Science Project Management. (n.d.). *CRISP-DM*. Retrieved 13 February 2021 from: https://www.datascience-pm.com/crisp-dm-2/
11. Downey, L. (n.d.). *Efficient Market Hypothesis (EMH).* Investopedia. Retrieved 5 May 2022 from: https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp
12. Gandhmal, Dattatray P. & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. *Computer Science Review,* (34), 100190.
13. GeeksforGeeks. (n.d.). *Python | Sentiment Analysis using VADER*. Retrieved 20 April 2022 from: https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/
14. Grigoryan, H. (2015). Stock market prediction using ANN Case study of Talit, Nasdaq OMX, Baltic Stock. *Database Systems Journal*, 2(6), 14-23.
15. Hladik, L. (n.d.). *Methods of Stock Market Prediction*. Muse – Union College Blogging. Retrieved 15 December 2021 from: https://muse.union.edu/2019capstone-hladikl/methods-of-stock-market-prediction-2/
16. I Know First. (2021). *Stock Market Forecast: Chaos Theory Revealing How the Market Works*. Retrieved 15 November 2021 from: https://iknowfirst.com/stock_market_forecast_chaos_theory_revealing_how_the_stock_market_works
17. Idrees, S. M., Alam, M. A. & Agarwal, P. (2019). A Prediction Approach for Stock Market Volatility Based on Time Series Data. *Institute of Electrical and Electronic Engineers*, 7, 17287-17298.
18. Java T Point. (n.d.). *Types of Data Mining*. Retrieved 20 April 2022 from: https://www.javatpoint.com/types-of-data-mining
19. Jupyter. (n.d.). *About Us, Project Jupyter's origins and governance*. Retrieved 25 April 2022 from: https://jupyter.org/about
20. JustAnotherArchivist. (n.d.). *Snscrape*. Retrieved 25 May 2022 from https://github.com/JustAnotherArchivist/snscrape
21. Kapoor, N. (2021). *Types of Sentiment Analysis and Its Uses*. Medium. Retrieved 20 February 2022 from: https://medium.com/swlh/types-of-sentiment-analysis-and-its-uses-ad733535c895
22. Kara, Y., Boyacioglu, M. A. & Baykan, O. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The

sample of the Istanbul Stock Exchange. *Expert Systems with Applications,* 38(5), 5311-5319.
23. Kedar, S.V. & Kadam, S. (2021). Stock Market Increase and Decrease using Twitter Sentiment Analysis and ARIMA Model. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12, 146-161.
24. Kesavan, M., Karthiraman, J., Ebenezer, R. T. & Adhithyan, S. (2020). Stock Market Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data. *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 477-482.
25. Kim, S., & Kang, M. (2019). Financial series prediction using Attention LSTM. ArXiv, abs/1902.10877.
26. Koceska, N., & Koceski, S. (2014). Financial-Economic Time Series Modeling and Prediction Techniques–Review. *Journal of Applied Economics and Business*, 2(4), 28-33.
27. Laerd Statistics. (n.d.). *Pearson Product-Moment Correlation*. Retrieved 13 March 2022 from: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php
28. Liew, J. & Wang, G. Z. (2016). Twitter Sentiment and IPO Performance: A Cross-Sectional Examination. *The Journal of Portfolio Management*, 440, 129-135.
29. Mehta, Y., Malhar, A. & Shankarmani, R. (2021). Stock Price Prediction using Machine Learning and Sentiment Analysis. *2nd International Conference for Emerging Technology (INCET)*, 1-4.
30. Ming, F., Wong, F., Liu, Z. & Chiang, M. (2014). Stock market prediction from WSJ: Text mining via sparse matrix factorization. *2014 IEEE International Conference on Data Mining*, 430-439.
31. Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. & Anastasiu, D.C. (2019). Stock Price Prediction Using News Sentiment Analysis. *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, 205-208.
32. Mondal, P., Shit, L. & Goswami, S. (2014). Study of effectiveness of Time Series modeling (ARIMA) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, 4(2), 13-29.
33. Nasdaq. (n.d.). *Apple Inc. Common Stock*. Retreived 26 Feruary 2022 from: https://www.nasdaq.com/market-activity/stocks/aapl
34. Nguyen, T.H., Shirai, K. & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24), 9603-9611.
35. NLTK. (n.d.). *Natural Language Toolkit*. Retrieved 25 May 2022 from: https://www.nltk.org/
36. Pandas. (n.d.). *User Guide*. Retrieved 25 May 2022 from: https://pandas.pydata.org/docs/user_guide/index.html

37. Pano, T. & Kashef, R. (2020). A complete Vader-based sentiment analysis of Bitcoin (BTC) tweets during the ERA of COVID-19. *Big Data and Cognitive Computing*, 4(4), 1-17.
38. Prabhakaran, S. (2021). *ARIMA Model – Complete Guide to Time Series Forecasting in Python*. Machine Learning Plus. Retrieved 15 February 2022 from: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
39. Python. (n.d.). *The Python Standard Library*. Retrieved 25 May 2022 from: https://docs.python.org/3/library/
40. Rajbhoj, A. (2019). *ARIMA simplified*. Towards Data Science. Retrieved 15 August 2021 from: https://towardsdatascience.com/arima-simplified-b63315f27cbc
41. Rapidminer. (n.d.). *Operators*. Retrieved 20 April 2022 from: https://docs.rapidminer.com/latest/studio/operators/
42. Roul, A. (2021). *Sentiment Analysis – Lexicon Models vs Machine Learning*. Medium. Retrieved 13 March 2022 from: https://medium.com/nerd-for-tech/sentiment-analysis-lexicon-models-vs-machine-learning-b6e3af8fe746
43. Rozinajova, V., Ezzeddine, A.B., Lóderer, M., Loebl, J., Magyar, R. & Vrablecová, P. (2018). *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*. Academic Press.
44. Shah, P. (2020). *Sentiment Analysis using TextBlob*. Towards Data Science. Retrieved 13 March 2022 from: https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524
45. Sharma, S. (2021). *Sentiment Analysis Using the SentiWordNet Lexicon*. Medium. Retrieved 13 March 2022 from: *https://srish6.medium.com/sentiment-analysis-using-the-sentiwordnet-lexicon-1a3d8d856a10*
46. Smailovic, J., Grcar, M., Lavrac, N., & Znidarsic, M. (2013). Predictive Sentiment Analysis of Tweets: A Stock Market Application. *CHI-KDD*.
47. Statista. (n.d.). *Leading countries based on number of Twitter users as of January 2022*. Retrieved 01 May 2022 from: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
48. Stedman, C. & Hughes, A. (n.d.). *Data Mining*. Tech Target. Retrieved 20 April 2022 from: https://www.techtarget.com/searchbusinessanalytics/definition/data-mining
49. Tang, X. & Yang, C. & Zhou, J. (2009). Stock Price Forecasting by Combining News Mining and Time Series Analysis. *Proceedings - 2009 IEEE/WIC/ACM International Conference on Web Intelligence*, WI 2009(1), 279-282.
50. Twitter. (2020). *#aapl*. Retrieved 20 December 2021 from: https://twitter.com/search?q=%23aapl&src=typed_query
51. Voigt, K. & O'Shea, A. (n.d.). *Stock Market Basics: What Beginner Investors Should Know*. Nerd Wallet. Retrieved 20 April 2020 from: https://www.nerdwallet.com/article/investing/stock-market-basics-everything-beginner-investors-know

52. Wang, Y. (2017). Stock market forecasting with financial micro-blog based on sentiment and time series analysis. *Journal of Shanghai Jiaotong University (Science)*, 22(2), 173-179.
53. Yahoo Finance. (n.d.). *Apple Inc (AAPL)*. Retrieved 20 December 2021 from: https://finance.yahoo.com/quote/AAPL/history?p=AAPL&guccounter=1

APPENDIX

Povzetek (Summary in Slovene language)

Medtem ko iščejo načine, kako pomnožiti svoje bogastvo, mnogi v nekaj vlagajo. Čeprav so zelo tvegane, so najbolj donosne naložbe delnice. Borza, verjetno eden najbolj priljubljenih trgov v današnjem času, temelji na nakupu delnic – deleža podjetja s pričakovanjem, da se bo njegova cena v določenem časovnem obdobju dvignila. Zato so borzne napovedi postale zelo priljubljene, zlasti z razvojem umetne inteligence in strojnega učenja. V zadnjem obdobju se pogosto pojavljata dva pristopa: analiza časovnih nizov (angl. Time Series Analysis, TSA) in analiza sentimenta (angl. Sentiment Analysis, SA). TSA pomeni iskanje vzorcev v zgodovinskem obnašanju vrednosti delnice, da bi se približali njeni prihodnji vrednosti. Drugi pristop, SA, je zelo priljubljen na področju obdelave naravnega jezika. Poskuša napovedati čustvo iz besedila. Namen moje magistrske naloge je implementirati in primerjati učinkovitost teh dveh pristopov ter raziskati, ali bi kombinacija obeh prinesla boljše rezultate. Da bi to naredila, implementiram ARIMA TSA, SA ki temelji na leksikonu, in kombinacijo teh dveh. Vsi trije modeli napovedujejo, če bo cena v naslednjih dneh višja ali nižja od trenutne cene za več kot 5 % za zaloge Apple Inc. Vse izvajam s programskim jezikom Python in Rapid Miner Studio, programsko platformo, ki se uporablja za rudarjenje podatkov. Za TSA uporabljam model ARIMA, ki je na voljo v Rapid Miner Studiu. Za SA uporabljam podatke twitterja za oceno dnevnega razpoloženja do trga. Z uporabo razlike v deležih pozitivnih in negativnih tvitov predvidevam, ali bo cena delnice narasla ali padla. V primerih velikih dnevnih nihanj pozitivnih in negativnih tvitov, torej če delež negativnih tvitov iz dneva v dan močno narašča, model zelo natančno napoveduje vedenje trga. Če pa ni večjih nihanj, bi model predvideval, da se cena ne bo spremenila, čeprav se v nekaterih primerih spremeni. Zato tega modela ni v samostojni uporabi. V 90 % opazovanih dni model TSA, ki temelji na ARIMA, pravilno napoveduje, ali bo cena delnice AAPL narasla ali padla. Čeprav TSA kaže zelo dobre rezultate, jih je še vedno mogoče izboljšati, in to storim tako, da dodam SA na vrh TSA, da ustvarim hibridni model obeh. Hibridni model se zanaša na TSA v času, ko ni velikih sprememb v splošnem razpoloženju. Če pa pride do velike spremembe, hibridni model prilagodi napoved TSA na podlagi SA. Končni rezultat resnično dokazuje, da je z združevanjem teh dveh pristopov mogoče bolje napovedati gibanje trga kot ju uporabljati samostojno. Hibridni model je natančno napovedal gibanje trga v 92 % opazovanih dni.

Magistarska naloga se osredotoča na napovedovanje, ali bo cena presegla ali bo pod določeno mejo, in ta relativno premalo raziskan pristop je ena izmed dodanih vrednosti mojega dela. Medtem ko je poudarek številnih dokumentov, ki temeljijo na SA, na celotnem feedu Twitterja, ta omejuje uporabnika na podlagi števila sledilcev, da bi se izognili tvitom, ki bi imeli nizek doseg in zato skoraj ne bi vplivali na gibanje cen.Tako SA kot TSA, torej hibridni model, je mogoče prilagoditi različnim trgom, saj se pragovi, ki so postavljeni in so osnova za rezultat napovedi, enostavno spremenijo glede na opazovani trg. Nazadnje, implementirane pristope je mogoče prilagoditi tako, da odražajo želeno donosnost naložbe.