Tampereen yliopisto

Tommi Mehtänen

KIRJOITA TÄHÄN TYÖSI PÄÄOTSIKKO

Kirjoita tähän työsi mahdollinen alaotsikko

Opinnäytetyön taso
Kirjoita tähän tiedekunnan nimi
Tarkastaja:
Tarkastaja:
Kuukausi Vuosi

TIIVISTELMÄ

Tekijän nimi : Opinnäytetyön otsikko
Opinnäytetyön taso
Tampereen yliopisto
Tutkinto-ohjelma
Kuukausi Vuosi

Avainsanat: Tiivistelmä-tekstin jälkeen.

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin Originality Check -ohjelmalla.

ABSTRACT

Author : Title
Thesis type
Tampere University
Degree Programme
Month Year

The abstract is a concise 1-page description of the work: what was the problem, what was done, and what are the results. Do not include charts or tables in the abstract.

Put the abstract in the primary language of your thesis first and then the translation (when that is needed).

This document template has two text styles for abstract. BibInfo is for bibliographical information above whereas the rest uses the style Abstract, which has line spacing of 1.0. The style Heading (no number) is used in the frontmatter before actual text and it makes the necessary preceding page break. Similar style is used in the bibliography with slightly different name in order to include it in the table of contents. The title page must end with Section Break to get pages numbered correctly. Moreover, the header on this page turns off the setting Link to Previous and formats the page numbers to Start at 1 (instead of Continue).

Keywords: After Abstract-text

The originality of this thesis has been checked using

In this thesis artificial intelligence tools and their use cases are described below:

OpenAI Chatgpt3.5:

the Turnitin Originality Check service.

ALKUSANAT

Tampereella, 15.2.2019

Päivittäjä

TABLE OF CONTENTS

KUVALUETTELO

Tämä luettelo on vapaaehtoinen. Kuvaluettelo lisätään *References* > *insert Table of Figures* ja sieltä *Options…* > *Build table of figures based on* > *Style:Figure Caption*. Myös taulukkoluettelon saa samasta kohdasta, kun valitsee viimeisestä kohdasta tyylin *Table Caption*.

LYHENTEET JA MERKINNÄT

SVM : Support vector machine
DSRM : Design science research model
.

1. INTRODUCTION

Predicting stock markets have been topic of interest for economics and researchers for a long time. Due to digitalization, amount of available information has increased. Users of social media platforms have increased so therefore impact of social media have also increased. Therefore, it has become interesting source of public data. There was estimated of 5.3 billion internet users and 4.95 billion social media users (Statista, 2023).

There are many different factors that effect on stock prices. There are technical and nontechnical factors. Social media platforms offer investors a place to collaborate and discuss stocks. There are indicators that sentiments of these discussions effect on stock markets.

1.1 Background

Predicting stock markets have been topic of interest for economics and researchers for a long time. Due to digitalization, amount of available information has increased. Users of social media platforms have increased so therefore impact of social media have also increased. Therefore, it has become interesting source of public data. There was estimated of 5.3 billion internet users and 4.95 billion social media users (Statista, 2023).

There are many varied factors that effect on stock prices. There are technical and nontechnical factors. Social media platforms offer investors a place to collaborate and discuss stocks. There are indicators that sentiments of these discussions effect on stock markets.

Data analytics and big data have become increasingly interesting topics when it comes to decision making. With a wide range of data sources social media has become more important.

There are many popular social media websites that has discussion about stocks like X, Facebook, Reddit and so on. When it comes to Finnish platforms Inderes forum is quite popular among investors.

1.2 Objectives and research question

Purpose of this research is to study Finnish stock markets with the lens of social media. Goal of this research is to come up with new kind of way to do social media data analytics on Finnish stock markets area. The main problem that was identified was that there is not always API to get clean and easy to use data from. In some social media platforms(X) there are API available, but it costs. Objective of this study was to design, develop and evaluate practical solution to overcome this problem.

The main research question:

How can Finnish stock markets be studied through lens of social media?

Besides the problem with collecting data, there were other practical problem identified. The problem was lack of tags or any type of labelling on forum discussions. Because of that identifying what kind of entity is in question in discussion. For example, for a computer program it is not that trivial to identify if in discussion “Nokia” means the company or location. The one side goal to reach solution to original problem was to identify forum posts that discuss on specific companies accurately.

1.3 Scope of this study

In this thesis the focus is only on Finnish stock markets and Finnish forums. Popular Finnish forum among investors is called Inderes. X is also very popular social media platform but due to its current prices it is left out of the scope. There are also several studies already made from X. There are also several other forums where Finnish investors discuss like Reddit, Vauva.fi and Ylilauta but they are not as active as Inderes so that is one reason why Inderes was selected. It seems like Inderes is the most active social media platform when it comes to Finnish investors.

In the beginning 55 largest Finnish companies were chosen to focus on but afterwards amount was filtered into top 43 companies because of some of the biggest firms were just listed recently so there were no a lot of data available. One of those companies was Mandatum Oyj which was listed in October 2023.

Also, timeframe of this research was limited to beginning in the January first, 2019, till end of December 2023. When deciding timeframe, the data availability in Inderes was a restriction. There were data available from 2018 until today. To have comprehensive and clear dataset that timeframe was selected.

1.4 Structure of the study

In introduction background of the study, research problem and scope of the study was introduced. In this section structure of the study is presented.

This thesis consists of following parts:

- Introduction
- Literature review
- Research methodology
- Artifact design and evaluation
- Conclusion and discussion

Literature review part of this study was about getting familiar with the research problem. The goal was to gather knowledge of the topic through literature. First, impact of social media discussion on financial markets is being presented.

The reason why this thesis was constructed that way was the choice on using design science research methodology as a way of conducting this study. After that data analysis techniques in financial forecasting with a focus on sentiment analysis is being introduced. These parts are essential to build strong understanding of the topic, what is already done and to identify the gaps in this research area.

After literature review research methodological decisions are being presented. In this chapter framework used in this work is also being presented.

2. LITERATURE REVIEW

In this chapter existing literature is being studied, introduced, and analysed. First, we look at the relationship between social media and stock markets. Then we look at the existing analysis techniques when it comes to stock predictions with a focus on sentiment of the social media posts. After that web scraping for data collecting is being presented.

2.1 The growing impact of social media discussion on financial markets

Forecasting stock markets has been a significant focus for researchers, employing various predictive models to decipher market movements. Charles H. Dow's theory, emphasizing human psychology's impact on market prices, underlines the importance of psychological factors in market predictions. This perspective suggests that market participants often expect current conditions to persist indefinitely, affecting their investment decisions (Chopra & Sharma 2021). Moreover, Zheng & Chen (2013) explored how supply and demand, influenced by long-term fundamentals and investor sentiment, play crucial roles in stock price determination. This highlights the multifaceted nature of stock market dynamics, where predictive models strive to unravel the complex interplay of various influencing factors.

Chen et al., (2019) demonstrated the potential of machine learning in financial forecasting by using support vector machines (SVM) to predict stock movements in China's markets with an accuracy of 88.16%, showcasing the efficacy of modern computational techniques over traditional methods. Similarly, Bollen et al. (2011) ventured into the realm of social media analytics, utilizing Twitter data to forecast the Dow Jones Industrial Average's daily changes, achieving an 86.7% accuracy rate. This study bridges the gap between predictive models and the burgeoning field of social media impact, highlighting the relevance of public sentiment in financial forecasts.

According to Piñeiro-Chousa et al. (2017) social media influence the stock market. Also, Antweiler & Frank (2004) found that internet stock messages help with predicting market volatility. Effect on stock returns was found to be statistically significant but economically small.

In 2022 there was a study made where they attempted to study impact of social media on stock markets. They focused on stock market of Odisha which is in India. They found out that social media has a vital impact on the investors in stock market. Relationship

between social media and investors of stock market was significant. (Patra et al., 2022) Although this study focused only on Odisha the findings backup this work.

When it comes to predicting closing price of the following day of a stock technical analysis, news articles and Twitter were used together and modelled using time series mining techniques, results were promising. There was a relation between price data and textual data. (Kollintza-Kyriakoulia et al., 2018)

In 2015 Twitter data was used to analyse the impact of tweets on Microsoft and Apple stocks. They found out that if the news in Twitter about Microsoft were positive it had 1 % significance level on stock return with one day delay. But when news was negative it had 5 % of significance level on Microsoft stocks. (Vojtěch et al., 2015) Based on these findings we could argue that negative news has more effect.

There seems to be multiple different kinds of studies made about social media's impact on stock markets. Based on the literature it is safe to say that there is a relationship between social media and stock markets. Most of the studies used data sources like Twitter (Kollintza-Kyriakoulia et al., 2018; Vojtěch et al., 2015; Bollen et al., 2011), StockTwits.com (Piñeiro-Chousa et al., 2017), Facebook (Siikanen et al., 2018) or Yahoo! Finance and Raging Bull (Antweiler & Frank 2004). Based on these findings about data sources I could argue that retrieving and analysing data from Inderes is new thing in this research area.

Only one of the studies focused on Finnish social media and stock markets (Siikanen et al., 2018). Other geographically limited studies were study made by Chen et al., (2019), which focused on China's stock markets. Patra et al., 2022 focused on India's stock markets and area called Odisha. Only one study was found that was specifically about Finnish social media and stock markets (Siikanen et al., 2018). This is one of the gaps in the research area that this study tries to fill in.

Most of the studies focused on the relationship between social media and stock markets or predicting stock movements rather than trying to answer the question on how this kind of study should be done. That is the main gap this study is attempting to fill.

Methods to collect data from desired social media platforms included tools like: Social Data Analytics Tool (Siikanen et al., 2018), crawler (Kollintza-Kyriakoulia et al., 2018), Twitter API (Vojtěch et al., 2015) or by using speicalized software (Antweiler & Frank 2004). There was no detailed information about collecting the data. It seemed to have been trivial case.

2.2 Data analysis techniques in financial forecasting with a focus on sentiment analysis

The advent of social media has introduced a novel dimension to financial market analysis, with several studies examining its influence on stock prices and investor behaviour. Eierle et al. (2022) identified a strong correlation between social media sentiment and short-term stock returns, particularly in stocks with negatively adjusted sentiments. This suggests that social media platforms can significantly affect market perceptions and movements. Pan (2020) further emphasized the role of investor sentiment, propagated through social media, in inflating market bubbles, indicating the profound psychological impact of online discourse on financial decisions.

Research by Siikanen et al. (2018) highlighted the differential impact of Facebook data on investor groups, where passive households and non-profit organizations were swayed by social media posts, unlike more sophisticated investors. This dichotomy underscores the nuanced effect of social media across different investor classes, pointing to the variegated influence of online sentiment on market participation.

Sentiment analysis has emerged as a pivotal tool for gauging the pulse of social media and its impact on markets. Li et al. (2017) leveraged a vast dataset of tweets to predict the stock movements of 30 companies with remarkable accuracy, using the SMeDA-SA technique. This underscores the potential of sentiment analysis in extracting actionable insights from social media chatter. Tarsi et al. (2023) further validated the efficacy of sentiment analysis by employing LSTM models to predict Amazon's stock prices based on Twitter and Yahoo data, finding a significant correlation between social media sentiment and stock price movements.

Koukaras et al. (2021) introduced the PageRank approach to refine sentiment analysis by accounting for the importance of webpages, showcasing the methodological diversity in sentiment analysis applications. These studies collectively highlight the advancements in sentiment analysis techniques, from traditional models like SVM to cutting-edge deep learning and NLP techniques, underscoring their critical role in financial market analysis.

Sentiment plays a huge role in the relationship between social networks sentiment and investors' decisions (Piñeiro-Chousa et al., 2017). Especially disagreement among the posted messages predicted increase in trading volume (Antweiler & Frank 2004). In this thesis, sentiment of the posts plays also important role. Based on the findings from the literature it is reasonable to be especially interested in sentiment analysis.

When it comes to analysing sentiments on a sentence level, there are many different natural language processing techniques to make predictions better. These techniques

include tokenization, lemmatization, stemming, filtering and so on. There are many different techniques to analyse sentiments of a text data but most of them are based on lexicon-based method or machine-learning methods. Recursive Neural Tensor Network is more complicated than an AFINN-model, but the performance was not better. (Zou 2019) This comparison between those two models were made for English language text. In this thesis we focus on Finnish language text so therefore that comparison might not be that suitable for this use case.

Because of the Finnish language typical sentiment analysis techniques might not perform well. Therefore, we need to look at more specific techniques tailored for Finnish language. Traditionally machine learning models like Support Vector Machine and naïve Bayes classifier have been sued for sentiment analysis but recently new deep learning techniques have emerged, and they look promising for sentiment analysis (Nukkarinen 2018)

There are several different options when it comes to Finnish language sentiment analysis. In this thesis sentiment analysis was only part of feature engineering. Toivanen et al 2022 states that when it comes to Finnish language sentiment analysis out of all BERT models FinBert performed the best. Therefore, in this thesis Finbert-model is being used in feature engineering.

2.3 Web scraping for data collection

As discussed in the section 2.1 there is different ways to collect data from internet. Typical way is using APIs or building your own way. In this case when main data source was chosen to be Inderes it turned out that there is no public API available. So that's why crawling through the website and scraping the needed content was decided to be a way to gather data.

Due to rise of internet users, there is also rise on available data on the internet. Web scraping is technique used for collecting data programmatically from a website. Web scraping has become a useful tool to make a sense of the information available online. (Jarmul & Lawson 2017)

Web scraping is a method used to collect data from a website without human manually collecting the data. "Or, in other words: instead of a human end user clicking away in a

web browser and copy-pasting interesting parts into, say, a spreadsheet, web scraping offloads this task to a computer program that can execute it much faster, and more correctly, than a human can." (vanden Broucke and Baesens, 2018). Web scraping makes it much easier to collect large amounts of data compared to manually collecting it.

Web scraping is recognized as a valuable tool for accessing online data in more efficient way than traditional data collection ways. When it comes to web scraping there is a need to consider ethical and legal things. Ethical guidelines for using online data are not clearly defined and continue to change as digital data collection techniques change. (Luscombe et al. 2022)

There are different tools for collecting data from static websites than from JavaScript-dependent websites (Jarmul & Lawson 2017). When it comes to Inderes website it was discovered that it is JavaScript-dependent website so that meant that typical web scraping techniques didn't work. Therefore, there was a need to choose from different techniques to successfully get the data. Two ways were considered which were PyQt and Selenium based on the book created by Jarmul & Lawson (2017).

Web data offers new way for researchers and practitioners to gain insight of a phenomenon. However, the novelty of web scraping means that legality and ethics are not yet well-defined, and they are considered of a grey area. (Krotov et al. 2020)

There are different technical solutions that a website provider can make to protect their website. These solutions can be banning an IP-address, Rate-limiting IP requests and other technical solutions. Many websites have terms and conditions that prohibit automated data collection. (Luscombe et al. 2022) In this research Inderes page was used and there was no prohibiting automated web collecting in terms and conditions, when data you are collecting is not used in a commercial way (Inderes 2024). So, if this web scraper is not used in commercial way there should not be legal issues.

Ethical considerations need to be considered. What this means in practice is the risk of sending too many requests to a website and therefore the traffic to the website is being overloaded. "The most obvious harm comes in the form of an unintended "denial of service "(DoS attack, which occurs when the high frequency and duration of a researcher's scraping algorithm overwhelms a website's server." (Luscombe et al. 2022) In this work only necessary requests were sent to the server and there was significant amount of time between them.

With ethical considerations there comes the question of what publicly available data is. It is not often that clear to say what data is public and what is not. (Ravn et al 2020) If we think about platforms like Facebook or Instagram the publicity of the data can be unclear.

That is because you can restrict availability of your posts for example to being you're your friends or people who follow you on that platform. When it comes to Inderes discussion forum it seems to be quite clear that the data of discussions is publicly available. There is no need for registration or any permission to enter the discussions.

3. RESEARCH METHODOLOGY

In this chapter all research methodological decisions are introduced and justified.

3.1 Research design

When doing research there are many choices researcher can do when answering research question or multiple research questions. In this work all research methodological decisions are shown in figure 1.

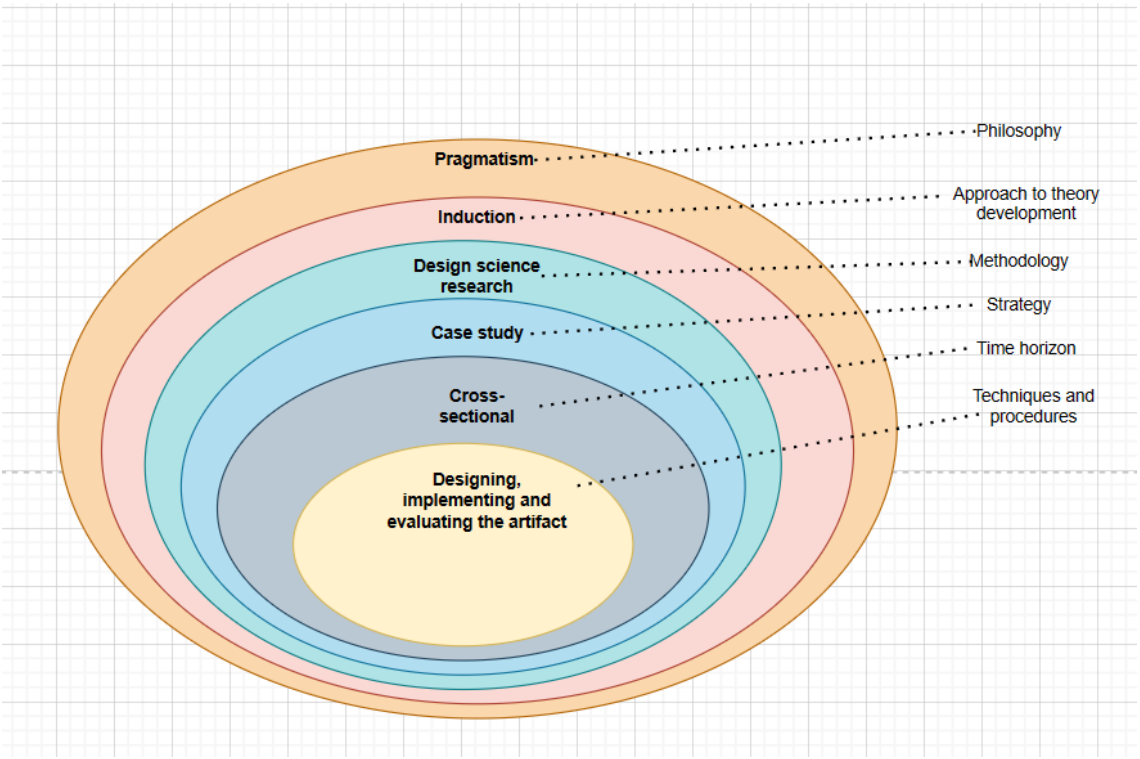


Figure 1. Research methodology. Modified from Saunders, Lewis and Tornhill (2019)

The first section of the onion model in figure 1 starts with deciding research philosophy. Research philosophy was chosen to be pragmatism because pragmatism does not focus that much on new theories, concepts, or ideas but it focuses on practical solutions for problems (Saunders et al. 2019).

Next section on the onion model presented in a figure 1 was deciding which approach to theory development would be suitable for this type of research. Based on Saunders et al. (2019) inductive approach is suitable if you are collecting data to explore a phenomenon and you are possibly making new theory or framework.

Choosing the methodology for answering the research question design science research was chosen. Since the nature of the research question and problem it was logical to choose design science research for this kind of research. Evaluation of the artifact had features of qualitative and quantitative research. Qualitative features were evaluating usability and readability of the artifact. Quantitative features in the evaluation were about amount of data provided for analytics and efficiency of the artifact. Since the usability of the artifact was recognized as a key feature on evaluation, it was logical to also focus on efficiency of the artifact because it affects directly on usability. Mixed methods are used for enhancing understanding, improve generalisability and to provide different perspectives (Saunders et al. 2019).

Next step was to choose research strategy. Based on preliminary exploration about the research problem this study was chosen to focus only on one social media platform Inderes since it seemed to be popular platform for Finnish investors. Therefore, the strategy to research on the problem was chosen to be case study. A case study strategy provides detailed insights by focusing on an intensive study of a specific phenomenon in a real-life context (Saunders et al. 2019).

The time horizon of this study is cross-sectional. Cross-sectional time horizon means that the research is tided to specific moment in time. It is like a “snapshot” taken at a particular time (Saunders et al. 2019). This study focused on the situation today and not the change of it in some time. There fore it was reasonable to choose cross-sectional time horizon.

Techniques and procedures to solve the research problem were chosen to be designing, implementing, and evaluating an artifact. The artifact collected the forum data from Inderes website and finance data from Yahoo Finance website. It attempted to show a way on how Finnish stock markets could be analysed using social media.

3.2 Design science research

In this study design science research was choice for research methodology. It was chosen because of the nature of the research question and research problem. This study focuses on designing, implementing, and evaluating a technical solution for a problem so design science research for research methodology seems appropriate.

There are many kinds of definitions about what design science research is. Hevner & Chatterjee (2010) defined it as follows: “Design science research is a research paradigm in which a designer answers questions relevant to human problems via the creation of

innovative artifacts, thereby contributing new knowledge to the body of scientific evidence. The designed artifacts are both useful and fundamental in understanding that problem". Creating an artifact sounds like a good way to approach this research question, because of the practical nature of the problem. When it comes to practical problems especially in the field of information technology it is reasonable to approach the problem with the intent to make an artifact to solve the problem. In this thesis the focus being solving the problem on how Finnish stock markets should be analysed through the lens of social media, it sounds reasonable to design, develop, and evaluate an artefact to overcome this problem.

To approach the research problem systematically and comprehensively DSRM process model was chosen as a guide for this research. To be able to make a choice on which DSRM process model should be used different models were compared.

According to Peffers et al. 2007 design science research consists of six steps. These steps consist of identifying the problem and importance of solving that problem, defining objectives of a solution, designing, and developing artifact, demonstration, evaluation, and communication. First four steps include entry points for research. This process model is shown in Figure 1. This study was made with focus on designing and developing artefact.

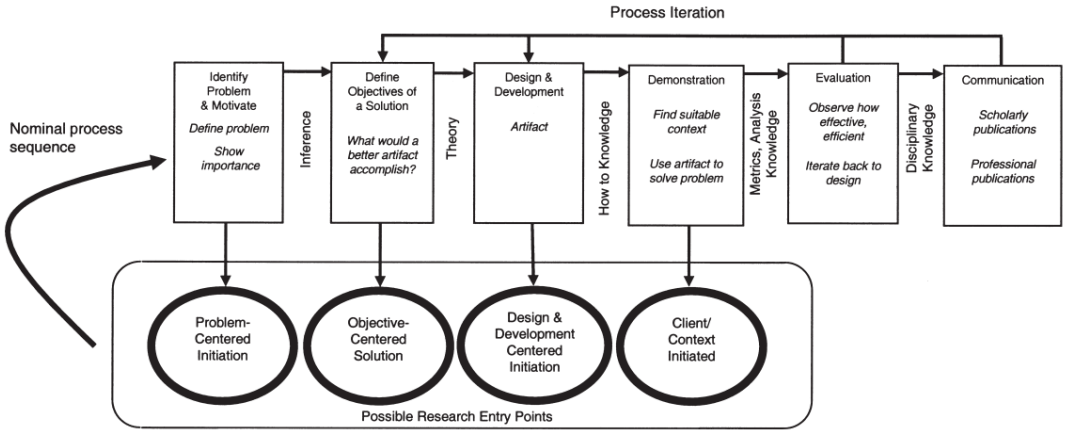


Figure 2. DSMR Process model (Peffers et al 2007)

This model represented in image 1 consist of six activities. First activity aims to identify the problem that is worth doing research on and motivate on why the problem is worth solving. In this activity you are supposed to justify your solution. (Peffers et., al 2007)

The second activity is about defining the objectives for a solution. Objectives can be quantitative or qualitative depending on the problem specification. This activity should have knowledge of current solutions and their efficiency. (Peffers et., al 2007)

The third activity is where artifact is being designed and developed. These artifacts can be constructs, models, methods, instantiations, or any designed object that is a contribution for the research. (Peffers et., al 2007)

The fourth activity is demonstration. In this activity you are supposed to show that your artifact works properly and solves one or more of the previously defined problems. The demonstration can be experiment, simulation, case study, proof, or other activity. (Peffers et., al 2007)

The fifth activity is evaluation. In this activity you are supposed to evaluate your artifact on how well it solves your previously defined problem. In this activity you can compare your objectives to results from your artifact. This activity can be performed in many forms including simulation, satisfaction surveys, client feedback, items produced or something else. (Peffers et., al 2007)

The last and sixth activity is proposed by Hevner & Chatterjee (2010), and it includes communication Chatterjee of your results. Usually, this activity is a research paper published. To communicate results of the study researchers can use this process to organize the structure of their paper. (Peffers et., al 2007) This process is used for a structure of this thesis as well.

Offerman et al 2009 compared different DSMR processes and introduced their own version of it. In table 1 this comparison is presented.

	Peffers et al. 2008 [52]	**Takeda et al. 1990 [60]**	**Nunamaker et al. 1991 [47]**	**March and Smith 1995 [42]**	**Vaishnavi and Keuchler 2004/5 [63]**	**Process presented here**
Problem identification	• Problem identification and motivation • Define the objectives for a solution	• Enumeration of problems	• Construct a Conceptual Framework		• Awareness of Problem	• Identify problem • Literature research • Expert interviews • Pre-evaluate relevance
Solution design	• Design and development	• Suggestion • Development	• Develop a System Architecture • Analyze & Design the System • Build the System	• Build	• Suggestion • Development	• Design artefact • Literature research
Evaluation	• Demonstration • Evaluation	• Evaluation to confirm the solution • Decision on a solution to be adopted	• Observe & Evaluate the System	• Evaluate	• Evaluation • Conclusion	• Refine hypothesis • Expert survey • Laboratory experiment • Case study / action research • Summarise results

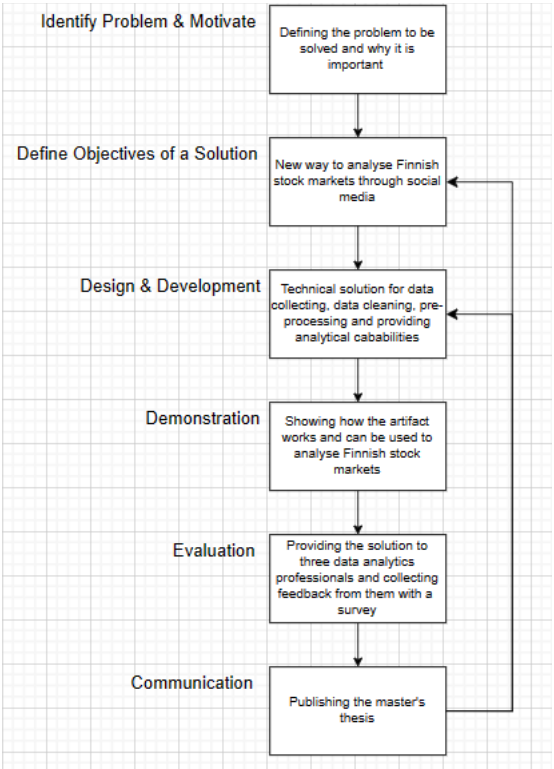
After reflecting research problem on the table 1, Peffers et al 2008 process model seems to be fitting this situation the best.

3.3 Artifact design and development process

To predict stock movements there are several options to choose from. Which model is best for your case depends on situation. To try to predict stock movements in there are indicators that support vector machine works best. At least it worked best in a study where they tried to predict Shanghai Composite Index based on social media platform for China's financial community data with the accuracy of 88.16%. Based on that SVM is also used in this thesis.

Forum discussion's comments were analyzed with sentiment analysis. Because of Finnish language FinBERT-model was chosen for technique. FinBERT-model was recognized to be most effective among single BERt models in Finnish language sentiment analysis (Toivanen et al. 2022).

Based on comparison of DSMR process models made by Offerman et al 2009 in table 1, one that was most suitable for this study was chosen to be process model made by Peffers et al 2007. Below in image 2 there is a more specific view of this research process.



The development of the artifact was done in two iterations. After the first iteration the artifact was evaluated through having a conversation and asking feedback from data-oriented professionals. After the first iteration it was discovered that there could be an improvement in data collecting phase from the Inderes forum which would improve the amount data of specifically related to stock markets. In practice this meant changing the starting point of the scraper from https://keskustelut.inderes.fi/ in to https://keskustelut.inderes.fi/c/osakkeet/17. This way there would be more comments in a dataset in the end of the process that would discuss specific stocks of interest. Implementing this change meant refactoring the scraper code because the website structure had changed and the scraper code that previously had worked did not work like before. For example, the number of visits on the discussion were not anymore always available in the same html-element as before.

Main challenges during the development of the artifact were about long execution times of code. To overcome this issue the software was divided into files so that the end user would not have to run the time-consuming parts. In this way the end user experience would be a little better.

After the first iteration of development process, it was decided that number of links and the title of the discussion were not important because in this case the point of interest was more on the comments than the discussion itself.

4. ARTIFACT DESIGN & EVALUATION

4.1 Artifact design

To collect data for this work there were two main data sources. Inderes forum was focus and topic of interest in this work. Other data source was Yahoo Finance. Data from these two sources were used to create an artifact for financial forecasting based on social media. The goal for the artifact was to be able to analyse financial markets with social media data combined with financial data. Due to the nature of this project, Jupyter Notebook was chosen to a tool for development and testing. Anaconda distribution was as an environment used for managing packages and libraries. Jupyter notebook was decided to be a tool for developing the artifact. Using normal python files to run the code instead of Jupyter Notebook could have reduced the execution time. To keep the artifact as transparent as possible Jupyter Notebook was chosen. It has a feature to run the code cell by cell in smaller code blocks. That feature and print statements help with transparency on what the code does in which parts.

The artifact was designed to be organized in three different files to help with managing the program. First file had a first part of the program, and it collected all needed data from the Inderes website in two parts and saved the data into two different files. This was because of possible errors happening during the run of the code. Even tough data collecting was an essential part of the project it was excluded from the evaluation of the artifact. Here is some reasoning behind that decision. Running the data collection code took extremely long, approximately two days, because of the dynamic structure of the website which prevented the use of typical scraper-libraries like Beautiful Soup or Scrapy. So, there was a need to use different tools to accomplish this html collecting and parsing task. Fitting library for this task was chosen to be Selenium. Maintaining the code for data collecting was not easy, because the structure of the website can change and break the code. Also, one thing making maintaining of the code difficult was that the code uses Selenium-library to get the desired data from the website and it is dependable on Google Chrome version and Chrome driver version. They need to be compatible. Therefore, decision was made to focus the evaluation of the artifact more on other parts of the solution. However, to be as transparent as possible, the data collecting code is available on the GitHub page of this project.

Second part of the program opened the two csv-files saved earlier in the first part of the program. This part of the program merged the datasets together. After that columns' datatypes were defined….

Another dataset was gathered from Yahoo Finance. One challenge here was that all the top 55 biggest Finnish companies have not been listed for a long time. For example, Mandatum was listed in October second, 2023.To overcome this challenge only companies which have been listed since beginning of 2019 until end of 2023 were filtered into the data frame. In the end dataset contains 41 companies.

4.2 Collecting data

Social media data was collected from Inderes forum and finance data was collected from Yahoo! Finance website. Yahoo! Finance offered an API to get the data from. Data from Inderes forum had to be scraped to obtain it.

4.2.1 Collecting the data from Yahoo Finance

Yahoo! Finance offers an API to retrieve data from. This API was used to obtain the data needed. In this work daily stock movements were not the case of interest so that's why finance data was processed to have stock price movement only on monthly aspect. Dataset was processed in a way that in the end finance dataset had three features. Below is a screenshot of selected features in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2460 entries, 0 to 2459
Data columns (total 3 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            2460 non-null   datetime64[ns]
 1   Ticker          2460 non-null   object
 2   Price Movement  2460 non-null   int64
dtypes: datetime64[ns](1), int64(1), object(1)
memory usage: 57.8+ KB
```

4.2.2 Collecting the data from Inderes forum

Since there were no public API available for data collection web scraper was build and used. Data from Inderes was collected using Jupyter-notebook with Python language. The idea was to first scrape a lot of data about discussions there. This decision leaves more options for end user. This way end user can decide which columns are useful.

Several decisions were made on what data to collect. The plan was to first collect as much data as possible and filter the data after collecting.

Scraping the data from Inderes was time consuming because Inderes discussion forum contains dynamic content so therefore in this study Python library Selenium was used to enter discussions and to scrape the html from discussions.

Due to the need for scraping as much data as possible and nature of the website, running the code to collect data of recent discussions on Inderes took two days to finish. After second development iteration the runtime was reduced to being around 7 hours. First data was collected in two parts and saved into csv-files, but after the first evaluation of the first version of the artifact complexity was reduced and data was collected all in one part and saved into a pickle file for later use.

Data about discussions were collected but so were data about individual comments in discussions. To be able to group comments together, ids of the comments were also collected to be able to later in the process map parts of each comment together. Elements that were collected from Inderes are presented in table 1.

Table 1. Data scraped from Inderes website.

Column	**Description**	**Type**
URL	URL address of the discussion	string
Created At	Timestamp of the creation of discussion	datetime
Last Reply	Timestamp of the last comment in the discussion	datetime
Visits	Number of visits	integer
Replies	Number of replies	integer
Users	How many unique users have commented	integer
Likes	Number of likes on the discussion	integer
Comments Details	All comments and their details.	List of dictionaries.

Collected dataset contained 789 discussions categorized as discussions about stocks. This was all the data that was available. Comments Details column contained of all the comments of a discussion and their details. Those details were the comments as a text, timestamp of comments, likes of comment, and id of the comment.

After all available data was collected and saved into a pandas data frame and saved as a pickle file for later usage. After the second iteration of development process execution time of this code took approximately 7 hours. At this point the timeframe of available data was from 2018 until the day of the execution of the code.

4.2.3 Data cleaning and preprocessing

Since the data collecting took 7 hours to execute preprocessing of the forum data was implemented on a separate Jupyter notebook file. This way after running the data collector script once there should not be needed to run it again.

After saving the data collected with the scraper it was time to clean and preprocess the data. Originally data was collected and saved in the format of one forum discussion per row. Since the point of interest was in comments individually rather than the whole conversation including the comments, the dataset was expanded into format of one comment per one row. In figure 3 there is a screenshot of what the forum data looked like after reading it into pandas dataframe.

```
First few rows of DataFrame:
                                                 URL   Created At  \
0  https://keskustelut.inderes.fi/t/onnistumiset-...   touko 2018
1  https://keskustelut.inderes.fi/t/macys-inc-tav...  maalis 2020
2  https://keskustelut.inderes.fi/t/suomalaisten-...    syys 2018
3  https://keskustelut.inderes.fi/t/nattopharma-j...   joulu 2020
4  https://keskustelut.inderes.fi/t/lindex-group-...   heinä 2021

    Last Reply  Visits Replies Users   Likes  \
0   joulu 2023  19,0 k      71    48     545
1    4. maalis  21,3 k     139    24     457
2    loka 2018   6,3 k      23     9      66
3  maalis 2021  15,4 k     138    35     546
4        2 pv   491 k   2,5 k   260  30,8 k

                                  Comments Details
0  [{'comment': 'Tässä ketjussa voi kehua omia on...
1  [{'comment': 'Macy’s Inc on perinteinen Yhdysv...
2  [{'comment': 'Harrastuksen vuoksi haluaisin py...
3  [{'comment': 'NattoPharma Corporate Video', 't...
4  [{'comment': 'Stockmann on kaikkien suomalaist...
```

Figure 3. Forum data before processing it.

In figure 3 is a sample of what the forum data looked like before preprocessing it. The data was in a format of one discussion per one row. The goal of preprocessing the data was to end up with a dataset in a format of one comment per one row and recognizing comments where certain companies are being discussed. The process of cleaning and preprocessing the data involved a lot of steps and turned out to be quite complex. For example, timestamps could be like “kesä 20”, “touko 19” meaning “June 2020” and “May 2019”. Timestamps were also sometimes like “2 pv” or “19 min” meaning “two days before the time of data collection” and 19 minutes before data collection. The artifact handled those timestamps into being in standard format for later use. Columns like “visits”, “replies”, “users”, and “post likes” sometimes contained data presented in a format like “2k”, “3m” and so on meaning 2000 and 3000000. The whole forum data cleaning and preprocessing is described in a Figure 4.

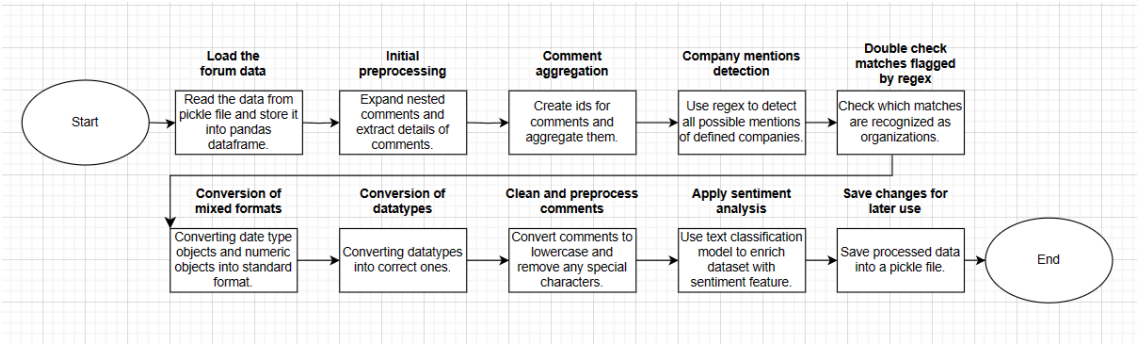


Figure 4. Overview of forum data cleaning and preprocessing.

Because Inderes website didn’t offer any labelling for comments identifying interesting comments had to be solved in some way. The forum data preprocessing stage involved accurately identifying comments discussing any of the leading 72 Finnish companies. Because the amount of forum data was limiting factor the goal was to first get a lot of data to be able to do more accurate analytics with it. Also available finance data was a restricting factor even tough soe forum data could be available. For example, there were some discussions about Mandatum Oyj but since it was listed in October 2023. To address this challenge, the selection of these companies was determined based on the market cap in January first, 2024.

Due to the nature of the data collected from the Inderes, there was a need to try and recognize and filter the comments that discussed about companies that we were interested in. As you can see in table 1 there are not any features that would tell what the comment is about. No hashtags or categories or any labelling. So therefore, another

solution for recognizing comments of interest had to be done. To overcome this challenge algorithm was designed and implemented and used to recognize interesting comments. First, we need to decide which companies we are interested in. In figure 5 there is an example of seven chosen companies and their regex patterns. The dictionary had 72 company names and their regex patterns. The patterns were decided to give matches easily so we would get all possible matches. In this way first there would be a lot of false positives but in the other hand the algorithm wouldn't miss any possible matches.

```
all_company_patterns = {
    "Nordea Bank Abp": r"(?i)\bnordea\w*",
    "Nokia Oyj": r"(?i)\bnokia\w*",
    "Mandatum Oyj": r"(?i)\bmandatum\w*",
    "Fortum Oyj": r"(?i)\bfortum\w*",
    "Sampo Oyj": r"(?i)\bsampo\w*",
    "Elisa Oyj": r"(?i)\belisa\w*",
    "Neste Oyj": r"(?i)\bneste\w*"
}
```

Figure 5. Example of defining companies of interest and their regex patterns.

After defining which companies are interesting and their corresponding regex patterns to match all possible mentions, model to categorize Finnish text was chosen to be finbert-ner model from kansallisarkisto. After checking if any of the regex patterns are detected in a comment, the comment is then analysed with the finbert-model to see if it recognizes it as an organization or not. All comments are looped through to enrich the dataset with the feature of companies mentioned in the comment. In the figure 6 there is the algorithm for detecting wanted companies.

```python
import re
from transformers import pipeline

# Initialize the NER model
model_checkpoint = "Kansallisarkisto/finbert-ner"
ner_classifier = pipeline("token-classification", model=model_checkpoint, aggregation_strategy="simple")

# Pre-compile regular expressions
compiled_patterns = {company_name: re.compile(pattern, re.IGNORECASE) for company_name, pattern in all_company_patterns.items()}

# Function to process comments
def process_comments(comment, compiled_patterns, ner_classifier):
    #print(f"Processing comment: {comment}")
    validated_matches = []

    # Loop through precompiled patterns and search for matches
    for company_name, pattern in compiled_patterns.items():

        match = pattern.search(comment)
        if match:
            match_text = match.group()

            # Use NER classifier directly inside the loop to minimize function calls
            ner_results = ner_classifier(match_text)
            is_org = any(entity_info['entity_group'] == 'ORG' for entity_info in ner_results)
            if is_org:
                validated_matches.append((match_text, company_name, is_org))
                print("Processed comment:", validated_matches)
    return validated_matches
# Apply processing to each comment in the DataFrame
print("Processing comments in DataFrame...")
data['Matched Companies Info'] = data['Comment'].apply(lambda x: process_comments(x, compiled_patterns, ner_classifier))
print("Processing complete.")
```

Figure 6. Implementation of the algorithm to identify comments discussing about any of the companies listed in an appendix 1.

Preprocessing the forum data and recognizing the companies of interest took around two hours and 23 minutes to complete. Because of that this part of the program was separated into a different file. With this decision the end user doesn't have to run the time-consuming part of the program if they don't want to, and they can focus more on the data analytics side.

After recognizing comments with mentions about any of the defined companies it was time to handle "Created At" and "Timestamp" columns. They contained spoken language descriptions of time. These columns were processed with code block presented in figure 7.

```
context_date = datetime.datetime.fromtimestamp(timestamp)

# Mapping for Finnish month abbreviations
month_mapping = {
    'tammi': '01', 'helmi': '02', 'maalis': '03', 'huhti': '04', 'touko': '05',
    'kesä': '06', 'heinä': '07', 'elo': '08', 'syys': '09', 'loka': '10',
    'marras': '11', 'joulu': '12',
    'tammikuuta': '01', 'helmikuuta': '02', 'maaliskuuta': '03', 'huhtikuuta': '04',
    'toukokuuta': '05', 'kesäkuuta': '06', 'heinäkuuta': '07', 'elokuuta': '08',
    'syyskuuta': '09', 'lokakuuta': '10', 'marraskuuta': '11', 'joulukuuta': '12'
}

# Function to convert Finnish date to standard format (with four-digit year)
def parse_created_at(date_str):
    parts = date_str.split(' ')
    if len(parts) == 2:
        month = parts[0].strip().lower()
        year = parts[1].strip()
        if month in month_mapping:
            return f"{year}-{month_mapping[month]}-01"
    return None

# Function to parse full timestamps like "25. toukokuuta 2022 5.49"
def parse_timestamp(date_str):
    parts = date_str.split(' ')
    if len(parts) == 4:
        day = parts[0].strip().rstrip('.')
        month = parts[1].strip().lower()
        year = parts[2].strip()
        time = parts[3].replace('.', ':')
        if month in month_mapping:
            return f"{year}-{month_mapping[month]}-{day.zfill(2)} {time}"
    return None

# Apply the parsing functions to your DataFrame columns
data['Created At'] = data['Created At'].apply(parse_created_at)
data['Timestamp'] = data['Timestamp'].apply(parse_timestamp)
```

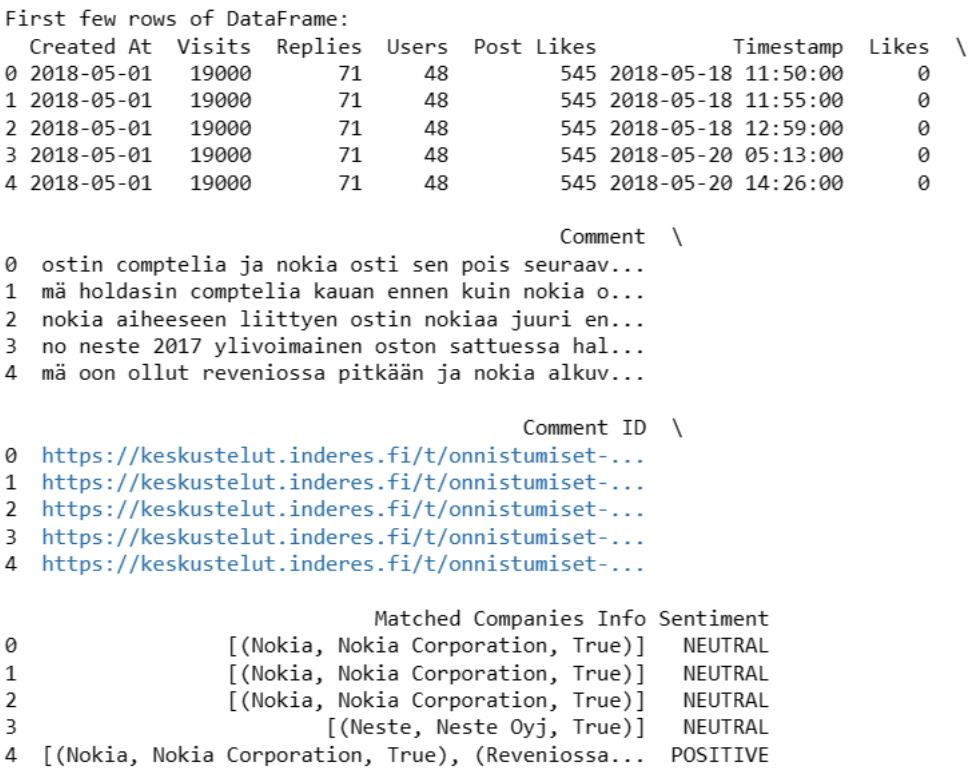
Figure 7. Code to handle columns containing timestamps.

After converting columns with timestamps in spoken language formats into ISO-standard format it was time to handle columns describing numbers into integers. Columns describing numeric data were also in unpopular formats and had to be changed into integers for later analysis.

Before calculating sentiments of the comments, they had to be pre-processed. This meant removing special characters and URLs. Text was also converted to lowercase.

Feedback gotten after the first iteration of the development process highlighted usability of the artifact to being bad because it was slow. To address this weakness, it was decided that calculating sentiments of the comments would be part of this file even tough it is categorized as part of feature engineering. This way the data-analysis notebook would be easier to use and would not take that much time to execute. More about sentiment analysis details in the next chapter.

Finally, after all the processing of the data including cleaning, recognition, parsing and calculating sentiments the data was stored into a pickle file for later use. Finally, this process presented in figure 4 ended up with data that is presented in figure 8.



```
First few rows of DataFrame:
  Created At  Visits  Replies  Users  Post Likes           Timestamp  Likes  \
0 2018-05-01   19000       71     48         545 2018-05-18 11:50:00      0
1 2018-05-01   19000       71     48         545 2018-05-18 11:55:00      0
2 2018-05-01   19000       71     48         545 2018-05-18 12:59:00      0
3 2018-05-01   19000       71     48         545 2018-05-20 05:13:00      0
4 2018-05-01   19000       71     48         545 2018-05-20 14:26:00      0

                                          Comment  \
0  ostin comptelia ja nokia osti sen pois seuraav...
1  mä holdasin comptelia kauan ennen kuin nokia o...
2  nokia aiheeseen liittyen ostin nokiaa juuri en...
3  no neste 2017 ylivoimainen oston sattuessa hal...
4  mä oon ollut reveniossa pitkään ja nokia alkuv...

                                       Comment ID  \
0  https://keskustelut.inderes.fi/t/onnistumiset-...
1  https://keskustelut.inderes.fi/t/onnistumiset-...
2  https://keskustelut.inderes.fi/t/onnistumiset-...
3  https://keskustelut.inderes.fi/t/onnistumiset-...
4  https://keskustelut.inderes.fi/t/onnistumiset-...

                           Matched Companies Info Sentiment
0                [(Nokia, Nokia Corporation, True)]   NEUTRAL
1                [(Nokia, Nokia Corporation, True)]   NEUTRAL
2                [(Nokia, Nokia Corporation, True)]   NEUTRAL
3                        [(Neste, Neste Oyj, True)]   NEUTRAL
4  [(Nokia, Nokia Corporation, True), (Reveniossa...  POSITIVE
```

Figure 8. Screenshot of the data after executing the preprocessing notebook.

If we compare the data before processing (figure 3) and after the processing (figure 8) we can see that there are a lot of changes happened. Now the data looks like it can be used for analytics. After all the processing we ended up with only 12578 rows of data.

4.2.4 Merging forum- and finance data

After preprocessing the forum data and collecting finance data it was time to read them on a separate notebook and merge those datasets together. Datasets were merged based on year, month, and company names in both datasets. By doing this inner join to merge datasets together new dataset contained forum data and finance data all in one data frame and ready to use. Merging the datasets into one dataset was decided to do because it would reduce complexity of the program and ease up the usage of the dataset.

4.2.5 Feature engineering

Sentiment analysis is categorized as part of feature engineering but since the aim is to provide efficient solution it was decided to keep as part of forum data preprocessing file. This way the end user would not have to wait for sentiments to be calculated but would get quicker into the analysis part. Rest of the feature engineering is in the explorative data analysis file.

4.3 Demonstration

To demonstrate developed artifact, it was decided to publish into GitHub. The link to the webpage of the artifact is included in the instructions in appendix a.

After collecting data from both sources and cleaning them. It was time for feature engineering. The goal was to think about features that are likely to have a correlation with stock price movement.

One main feature was to turn daily stock movements into monthly stock movements because monthly timeframe was selected. From daily closing and ending prices which were exact numbers, monthly price movement were derived in a way that 0 meant no changes, -1 meant price has gone down and 1 meant price has gone up.

Sentiment analysis of the comments was also chosen as an additional feature into the dataset. This was done with pretrained FinBERT-model because according to Toivanen et al 2022 it performed the best out of four BERT-models.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4314299

4.4 Evaluation

The evaluation was conducted with presentation of the artifact which leaded into discussions about how it could be better. Three data related professionals were asked to give feedback about the artifact. First feedback was conducted after the first iteration. Based on the feedback changes were developed to the artifact in the second development iteration.

The biggest challenge in developing an artifact was efficiency. After the first iteration based on the feedback runtime of the program was an issue. Because of the dynamic

structure of the website, Selenium was used instead of traditional web scraping libraries. This increased the runtime of the data scraper.

Second issue addressed in a development phase was that after the first iteration, the website's structure had changed so the scraper code had to be refactored to be able to collect data needed.

One challenge also was that there were no tags or any labelling when it came to comments. The data itself didn't contain the information about which stocks are they talking about. Therefore, own way to solve this problem had to be developed. This issue was tackled using regex for pattern recognition and NER-model to see if the matched pattern was recognized as an organization or no.

In the future there could be several improvements in the artifact. For example, in the data collection part also data from users could be collected. This is because different users have different influence even though they would be posting exactly same content. Someone who is very popular in the forum has more influence on their posts than someone who is not popular. Collecting data from users would enable more versatile analytics for end users.

5. CONCLUSIONS AND DISCUSSION

TODO:

Data anyst notebookki valmiiks.

Githubiin tavarat

Ohjedokumentti

Data cleaning and preprocessing jutskaa

Kuvat kuntoon

Demonstration kappaleeseen tavaraa.

Härpäkkeen ison kuvan muokkaus.

Käyttöliittymä jossa voi valita ajaako koko roskan alusta vai keskittyykö vain analyysi vaiheeseen.

REFERENCES

[1] Antweiler, W. & Frank, M. Z. (2004) Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of finance (New York)*. [Online] 59 (3), 1259–1294.

[2]

[3] Bollen, J. et al. (2011) Twitter mood predicts the stock market. Journal of computational science. [Online] 2 (1), 1–8.

[4] vanden Broucke, Seppe. & Baesens, Bart. (2018) Practical Web Scraping for Data Science Best Practices and Examples with Python. 1st ed. 2018. [Online]. Berkeley, CA: Apress.

[5] Eierle, B. et al. (2022) Does it really pay off for investors to consider information from social media? International review of financial analysis. [Online] 81102074-.

[6] Chen, S. et al. (2019) Predicting the Stock Price Movement by Social Media Analysis. Journal of data analysis and information processing. [Online] 7 (4), 295–305.

[7] Chopra, R. & Sharma, G. D. (2021) Application of artificial intelligence in stock market forecasting: A critique, review, and research agenda. Journal of risk and financial management. [Online] 14 (11), 1–34.

[8] Hevner, Alan. & Chatterjee, Samir. (2010) *Design Research in Information Systems Theory and Practice*. 1st ed. 2010. [Online]. New York, NY: Springer US.

[9] Jarmul, K. & Lawson, R. (2017) Python web scraping : fetching data from the web. Second edition. Birmingham: Packt.

[10] Kollintza-Kyriakoulia, F. et al. (2018) Measuring the impact of financial news and social media on stock market modeling using time series mining techniques. Algorithms. [Online] 11 (11), 181-.

[11] Koukaras, P. et al. (2021) “Predicting Stock Market Movements with Social Media and Machine Learning,” in International Conference on Web Information Systems and Technologies, WEBIST - Proceedings. 2021 pp. 436–443.

[12] Krotov, V. et al. (2020) Tutorial: Legality and ethics of web scraping. Communications of the Association for Information Systems. [Online] 47 (1), 539–563.

[13] Li, B. et al. (2017) Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Information systems (Oxford). [Online] 6981–92.

[14] Luscombe, A. et al. (2022) Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. Quality & quantity. [Online] 56 (3), 1023–1044.

[15] Nukarinen, V. (2018) Automated text sentiment analysis for Finnish language using deep learning.

[16]

[17] Offermann, P. et al. (2009) “Outline of a design science research process,” in Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, DESRIST ’09. [Online]. 2009 ACM. pp. 1–11.

[18] Pan, W.-F. (2020) Does Investor Sentiment Drive Stock Market Bubbles? Beware of Excessive Optimism. The journal of behavioral finance. [Online] 21 (1), 27–41.

[19] Patra, S. et al. (2022) Impact of Social Media on Stock Market - A Study on Odisha State. Shanlax International Journal of Management. [Online] 9 (4), 43–48.

[20] Peffers, K. et al. (2007) A Design Science Research Methodology for Information Systems Research. *Journal of management information systems*. [Online] 24 (3), 45–77.

[21]

[22] Petrosyan, A. (2024) Worldwide digital population 2024. Statista. Available at: https://www.statista.com/statistics/617136/digital-population-worldwide/ (Accessed:1.6.2024]).

[23] Piñeiro-Chousa, J. et al. (2017) Influence of Social Media over the Stock Market. *Psychology & marketing*. [Online] 34 (1), 101–108.

[24]

[25] Ravn, S. et al. (2020) What Is “Publicly Available Data”? Exploring Blurred Public–Private Boundaries and Ethical Practices Through a Case Study on Instagram. Journal of empirical research on human research ethics. [Online] 15 (1–2), 40–45.

[26] Saunders, M. N. K. et al. (2019) Research methods for business students. Eighth edition. Harlow, England: Pearson.

[27] Siikanen, M., Baltakys, K., Kanniainen, J., Vatrapu, R., Mukkamala, R., & Hussain, A. (2018). Facebook Drives Behavior of Passive Households in Stock Markets. Finance Research Letters, Forthcoming. Available at SSRN: https://ssrn.com/abstract=3134362 or http://dx.doi.org/10.2139/ssrn.3134362.

[28] Tarsi, M. et al. (2023) “Predicting stock price using LSTM and Social Media dataset,” in 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). [Online]. 2023 IEEE. pp. 1–4.

[29] Toivanen, I., Lindroos, J., Räsänen, V. & Taipale, S., 2022. Dealing with a small amount of data: developing Finnish sentiment analysis. In: 2022 BESC: 9th International Conference on Behavioural and Social Computing. [online] IEEE. Available at: https://doi.org/10.1109/besc57393.2022.9995536 [20.1.2024].

[30] Vojtěch Fiala et al. (2015) Impact of Social Media on the Stock Market: Evidence from Tweets. European Journal of Business Science and Technology. [Online] 1 (1), 24–35.

[31] Zheng, Xiaolian. & Chen, B. M. (2013) Stock Market Modeling and Forecasting A System Adaptation Approach. 1st ed. 2013. [Online]. London: Springer London.

[32] Zou, N. (2019) Text Analytics Methods for Sentence-level Sentiment Analysis.

[33] https://companiesmarketcap.com/finland/largest-companies-in-finland-by-market-cap/

APPENDIX A: LIST OF COMPANIES OF INTEREST

APPENDIX A: INSTRUCTIONS TO USE THE ARTIFACT