Norwegian University
of Life Sciences

Master's Thesis 2022 30 ECTS
Faculty of Science and Technology

Financial Sentiment Analysis of Quarterly Reports and Stock Performance

Michael Lindberg

Industrial Economics and Technology Management

Acknowledgments

My sincere gratefulness must be given to my engaged supervisors, Habib Ullah and Jesper Frausig, at NMBU for their guidance and presence. Their direct feedback and genuine interest in the research field have inspired me to find the right motivation for the thesis and helped me work in a creative way with a critical mindset. Also, I would like to express my gratitude to Habib, Jesper, and the other lecturers at NMBU for the inspirational courses.

Furthermore, I want to say thank you to my colleagues at the Norwegian Labor and Welfare Directorate for their support, engagement, and flexibility regarding my thesis work. It has been inspiring to discover the relevance of my thesis in my current job.

Finally, I would like to thank my family and friends who have always been supportive and interested in my work. I am extremely grateful to my parents for their inexhaustible support and encouragement throughout the work.

Norwegian University of Life Sciences

Ås, May 2022

Michael Lindberg

Abstract

This thesis aims to examine the use of financial sentiment analysis for quarterly reports published by companies listed on the Oslo Stock Exchange (OSE). Additionally, the intention of the study is to use methods from computer science to enable the transformation of financial reports, from the raw PDF format to the financial sentiment scores. Furthermore, this thesis aims to discuss the relationship between predicted financial sentiment and stock performance for chosen companies and industries. This thesis applies the famous and recently developed language model for financial sentiment analysis, FinBERT. The model is built upon a more general language model, BERT.

The motivation for the study is the increasing interest in machine learning and Natural Language Processing (NLP) for financial applications. Modern modeling techniques are allowing investors to make more informed decisions, and the rise of language modeling has made it possible to derive insight into the opinions of people through news and social networks. However, there are only a minority of studies investigating the language of quarterly reports.

Methodologically, quarterly reports from the first quarter of 2019 to the fourth quarter of 2021 are downloaded from the investor relations pages of the selected companies. The downloaded reports are the input of a data pipeline that extracts the text and predicts the financial sentiment using Python tools such as PDFMiner and the Transformers library. The predicted sentiment is then loaded into a pipeline for visualization and stock performance comparisons based on stock data downloaded with the yfinance open source tool.

The thesis concludes that extracting text from financial PDF files is feasible. Furthermore, the FinBERT model predicts the financial sentiment with a higher accuracy than the more general BERT model. However, the relationship between stock performance and predicted sentiment is not strong, despite individual differences. Additionally, the relationship is stronger for stock performance in the past. However, this thesis demonstrates the value of domain-specific NLP for applications in the financial industry.

Keywords – NLP, BERT, FinBERT, Sentiment Analysis, Decision-Making, Quarterly Report, Stock Prediction, Stock Performance, Pipeline

Contents

List of Figures

List of Tables

1 Introduction

This introductory chapter presents the broad background and motivation for the work presented in the thesis. The background section is focused on the application of language modeling in the financial industry, and the desired outcome is to give the reader a clear idea of the initial problem statement and the position of the thesis in the existing field of research. Furthermore, this chapter systematically presents the research goals and research questions in Section 1.2. Additionally, the section gives a brief explanation of the choice of data and methods for the initial problem statement, which are elaborated in Chapter 3 and 4. This chapter will clarify the importance of the work and the value provided in the field of financial sentiment analysis. In conclusion, an overview of the thesis structure is provided.

1.1 Background

The rational goal of an organization is to provide value. As hundreds of millions of events occur every day and are stored in large volumes [1], the focus on derived value from data is increasing [2]. Tools that interact with Big Data are of great interest for many organizations, as modern modeling technology enables prediction of behavior and events. Due to this, data are considered one of the most valuable commodities in automated systems for decision-making. A large number of organizations utilize these systems within different industries. Big data has become an important part of the financial industry [3], and financial practitioners use various data to accomplish predictive analyses and monitor the market. Additionally, the financial industry utilize the data and technology to enable efficient and accurate information processing. As a result of this, financial applications such as risk analysis, real-time analysis, fraud detection, algorithmic trading, and consumer analysis [2] have been born. However, a larger number of potential use cases are still in many ways unmapped ground due to the increasing amount of unstructured data, such as text data. On the other hand, the financial institutions and industry are discovering the value of textual data for applications such as chatbots for customer engagement and text analysis for market insight [4]. The extraction and structuring of unstructured financial text is a challenge, but an important step to enable the realization of profits from data.

Accurate text extraction and analysis can be a foundation for financial decision-making systems and deliver value for tasks such as investment decisions.

1.1.1 Natural Language Processing in Finance

The rise of modern modeling techniques descending from the field of Artificial Intelligence (AI), is resulting in a more innovative digital product development in the financial industry. Machine Learning (ML) and Deep Learning (DL) are of enormous interest, as the techniques enable financial organizations to perform predictive analyzes through a broad range of algorithms on the basis of a large set of data sources. Big data exists in various formats, and the potential outcome value of big data applications, such as actionable insight, would depend on a large number of factors, such as data quality, relevance, and the chosen modeling technique.

For financial practitioners, unstructured data such as news, websites, reports, research, and social media can be of interest for decision-making. Traditionally, it has been a challenge to do large-volume language modeling for domain-specific corpora. Financial professionals are reading and concluding based on their knowledge and experience, but new domain-specific language models equip practitioners with tools to improve oversight through large-scale objective evaluations [5].

Language modeling has become more efficient due to the latest advances in the field of Natural Language Processing (NLP) [6]. NLP, which is a subfield of linguistics, computer science, and AI, is gaining popularity in the financial industry [5], as the prediction accuracy of domain-specific language models is improved for financial applications. The latest advances in NLP, such as the availability of state-of-the-art language models like BERT [7], have opened a new research area for sectors such as marketing, finance, and economics.

1.1.2 Monitoring the Stock Market

Financial markets have a considerable impact on the economy of nations [8], as they are a pillar of the economy. Over the last few years, the attention to stock trading has increased. The application of ML techniques to the modeling of traditional financial data sources has been a common method to try to predict the stock market. Technical indicators, macroeconomic variables, and fundamental indicators have often been used to train ML models to predict future stock performance [9]. However, as interest in big data and ML techniques is increasing, the monitoring of the stock market with new analysis techniques is increasing in popularity. Modern stock trading involves a comprehensive use of technology. As traders demand information to maximize their profits, the race to obtain relevant information is of great interest [10]. Hence, it is desired to support financial analysts by making financial data accessible with the use of modern modeling techniques, such as financial language models.

As investors search for tools and modeling techniques to increase their profit and reduce risk, the stock market is known to be a difficult domain to predict. The Efficient Market Hypothesis (EMH) is a hypothesis that claims that the price of a stock reflects all available information. According to EMH, it is impossible to predict the market with 100% accuracy [11]. However, this has not ended the numerous attempts to maximize profit using a range of complicated modeling techniques and a wide range of data sources. Numerous studies have concluded that stock prediction is not a random walk and, to some extent, can be predicted [12, 13, 14]. The rise of NLP in finance makes stock monitoring through textual data possible and may equip financial decision-makers with actionable insight for investment decisions.

1.1.3 Financial Sentiment Analysis

In recent years, domain-specific NLP models have been built, such as the financial BERT model, FinBERT [15]. The FinBERT model offers powerful mechanisms to predict the financial tone of a text. This type of analysis is often referred to as sentiment analysis, which is one of the NLP tasks that recently achieved improved results [6]. Another study [10] emphasizes the potential for sentiment analysis, as the number of NLP algorithms for fast financial sentiment extraction is increasing. Obtaining more information in efficient ways is considered a competitive advantage for financial decision-making, such as stock trading [10]. Sentiment data such as financial news, reports, and tweets have been found to increase the accuracy of stock prediction models [16]. As investing has become more available and popular among non-professionals, the interest in human behavior and trends has increased. Especially, interest in the application of language modeling and sentiment analysis of tweets, news, and reports related to finance has increased.

Sentiment analysis is one of the main language modeling techniques applied for financial forecasting [17], and a study by Bollen et al. [18] found a strong correlation between the performance of a stock and the sentiment of people towards it. Applying sentiment analysis to news headlines alongside stock market data has shown that news has a measurable effect on the stock market [19]. Another study uses the BERT model for stock price prediction [20] along with technical indicators. Their research emphasizes that sentiment analysis impacts the market situation. Furthermore, trading sentiment portfolios outperform the benchmark index according to Bloomberg [15]. This is causing interest among practitioners for sentiment analysis in the financial industry and is suggested as future work in stock market research [19]. Monitoring the sentiment of markets through news, social networks, and financial reports can be used for investment decisions and other finance-related problems [15]. Billions of social media users and affordable computing power have made this resource-demanding analysis gain popularity.

As sentiment analysis of news and tweets has been the basis for stock prediction models [19], few studies apply sentiment analysis to financial reports published by listed companies. In a study by Petr Hájek, a strong relationship was reported between textual information extracted from annual reports and abnormal stock return [21]. However, the author suggests focusing on individual firms and industries in the future. In addition, another study shows that more frequent financial statements contribute to better predictions [22]. The intersection of these suggestions form the discovered gap in the research, which this thesis intends to fill. The use of quarterly reports for the purpose of measuring the relationship between financial sentiment and stock performance has not been attempted. Prediction of cash flow based on quarterly data has been attempted [22], but there are not many studies that apply sentiment analysis to published quarterly reports. As the language of financial reports is formalized, it is reasonable to believe that sentiment prediction with generalized language models, would not result in sensational sentiment scores. However, predicting the financial sentiment with FinBERT may reveal the true financial tone of a quarterly report. Furthermore, a study using BERT for stock prediction [23], suggests that future research should be based on longer texts such as call transcripts and financial reports. The hypothesis is that more information would lead to a more accurate stock prediction. The obligation of a quarterly report publication is to provide accurate information on the performance of a company. Hence, researching financial sentiment prediction of quarterly reports can add value to the field of financial analysis.

The researchers of the FinBERT model are suggesting utilization of the model for applications to finance-related outcomes such as stock returns, stock volatility, and corporate fraud [15]. According to a survey by Liapis, Karanikola, and Kotsiantis in November 2021 [24], FinBERT performed the best for sentiment analysis for financial time series forecasting. Using the FinBERT model for financial sentiment analysis of quarterly reports for specific firms may also contribute to this research. Demonstrating the use of domain-specific language models for the analysis of reports may contribute to the field of language processing in general. In addition, it demonstrates the potential of prediction of financial sentiment for better decision-making.

1.2 Research Goals and Questions

As mentioned above, few studies have been conducted on the financial sentiment of financial reports. Additionally, covering the gap in sentiment analysis of quarterly reports using the state-of-the-art FinBERT model may contribute to the research field of financial sentiment analysis. As it is suggested to focus on firm-specific documents [21], a primary research question can be formed for this thesis.

Does the FinBERT model reveal the financial tone of a quarterly report and demonstrate potential for financial decision-making for Norwegian technology, real estate and energy companies?

This study approaches existing research in a new way, applying state-of-the-art language models trained specifically to predict the financial tone of quarterly reports. The value of this thesis is to close the gap in the research, but also generalize the understanding of domain-specific language models. The pace of technology development is rapidly increasing, and therefore, a scope of opportunity for this exact research is created. This work focuses on modeling unstructured data, and the lack of research on sentiment analysis of quarterly reports and the few publications applying the FinBERT model for financial sentiment analysis form the narrowed scope of the study.

1.2.1 Research Questions

Subproblems have been derived from the main research question to give a clear idea of which specific questions this study intends to answer. The first question is a starting point for the work and is essential to answer the remaining research questions. A common method for validation of machine learning models and sentiment prediction is measuring the accuracy [25]. Human labeling can be performed to achieve a reliable evaluation of financial sentiment predictions. This way the accuracy can be computed based on domain knowledge and contextual understanding, and result in a trustworthy evaluation of financial sentiment prediction of quarterly reports.

1. Does the FinBERT model predict the financial sentiments of quarterly reports accurately and is the accuracy higher than for the general BERT model?

The second research question of the study concerns the future and past performance of stocks from the date a report is published. Previous research shows a strong relationship between stock performance and the sentiment predicted from different text sources [18, 19, 15]. Therefore, research on the relationship between stock performance and financial sentiment may add knowledge to the effect of financial reports on the equity market. As many studies use the one-day return of shares as a measure of stock performance [20], research can be extended to returns from longer periods of the future and past. Furthermore, the question involves research on the communication of companies based on the actual performance of the reported quarter. The researchers behind the FinBERT model recommend the detection of corporate fraud as a potential application of their model [15]. Therefore, it is interesting to study the performance of a company compared to the choice of language for the reported quarter. These two ideas form the scope of the second research question.

2. Is the predicted financial sentiment in quarterly reports strongly correlated with past and future movements in the equity market for a chosen stock?

Another question that the study wants to discuss is the difference in the correlation between asset-light and asset-heavy companies. Asset-light companies own fewer capital assets such as land, buildings, and machinery. On the other hand, asset-heavy companies own fixed assets, which are inevitable to generate income for the company. The categorizing of companies into asset strategies is not systematically stored financial data and has to be assessed by having company insight. A study by Zhou et al. [22] concludes that predicting the performance of companies based on quarterly data varies across the industry and suggests that the predicting gain is larger for asset-heavy industries. The study points out that investors in portfolios with a heavy weight in industries with stable cash flow can benefit from attention to the quarterly financial results. Furthermore, the same study suggests that future research should focus on analyzing individual companies on the basis of financial reporting. The familarity with industries and companies is important for the interpretation of the results.

3. Does the strength of the relationship between predicted sentiment and stock performance depend on the asset strategy for technology, real estate, and energy companies?

1.3 Desired Outcome and Methods

1.3.1 Desired Outcome

The initial purpose of the research is to develop knowledge of the sentiment analysis of financial reports, but also to provide a suggested workflow for financial analysts to use NLP for decision support. In addition, the thesis provides value to text extraction from PDF reports, which can be applied in many industries. The thesis adds an overall value to the task of deriving the financial sentiment of any financial report with the use of ETL pipeline structures, which are a set of processes to Extract, Transform, and Load (ETL) data to a desired destination. Furthermore, the study intends to demonstrate the use of state-of-the-art language model specifically pre-trained for financial applications. It is desired to compare the accuracy of the FinBERT model with the general BERT model for textual data related to finance. Furthermore, a desired outcome of the work is to add value to the analysis of the equity market using modern modeling techniques and non-traditional data sources. This includes discovering the potential of sentiment analysis related to the past and future performance of stocks across industries with different asset strategies. In other words, this research intends to investigate whether there exists any relationship between stock performance and the predicted financial sentiment of quarterly reports. In conclusion, the research field can utilize domain-specific NLP to gain insight into the communication of a company, which can be applied for various tasks such as detecting corporate fraud and misleading communication. Additionally, this thesis intends to demonstrate the value of domain-specific NLP for decision-making and oversight for financial analysts.

1.3.2 Data and methods

This thesis will focus on companies listed on the Oslo Stock Exchange (OSE) and apply text analysis to quarterly reports of companies. The chosen stocks are large-cap companies listed on the OSE and represent industries such as technology/telecommunications, real estate, and oil/gas. The reason for choosing large-cap companies is the expected availability and reliability of the reports of large companies. In addition, the chosen companies are well known, which makes evaluation of predicted sentiment an easier task. The reports are collected from the investor relations page of the selected companies. The time span for the reports downloaded is from the first quarter of 2019 to the fourth quarter of 2021. The hypothesis is that sentiment prediction with the FinBERT model can result in a high prediction accuracy of quarterly reports and potentially be related to past and future stock performance. The reason for choosing OSE and the mentioned industries is the knowledge of the author of the Norwegian stock market and the availability of data. This helps contextual understanding and manual labeling of sentences. Additionally, Norway has a transparent financial structure, and listed companies are obligated to report their results under strict regulations. This may improve data quality and the foundation for comparison. The data are described in more detail in Chapter 3.

Text extraction and analysis are performed using Python libraries and packages. In addition, Python is known to be an efficient programming language and is commonly used for NLP-related tasks. Furthermore, the intention of the study is to demonstrate a workflow in one programming language. This covers everything from the extraction of text from the PDF files to the visualization of stock returns with sentiment scores for each quarter. An important and time-consuming task of the study is to format the data and extract the text reasonably. Tools like PDFMiner and Pandas are important to solve this problem. Furthermore, plotly and yfinance help the comparison of stock performance and financial sentiments. The pre-trained FinBERT model is essential to execute the financial sentiment analysis and is downloaded and applied with the Transformers library in Python. The experience and knowledge of the methods are the main reasons for the choice. Additionally, Python is known as an efficient programming language and is commonly used for NLP-related tasks. More details about the methods and choices are provided in Chapter 4.

1.4 Thesis Structure

The thesis is structured with an explanation of the relevant theory in Chapter 2. The chapter presents main concepts relevant for the study, but also more technical components. Furthermore, Chapter 3 describes the chosen data and materials relevant to answer the initial research questions. The chapter explains how and why these are collected and prepared for analysis. The applied methods describe how the study is carried out and are explained by the structure and illustration of two pipelines in Chapter 4. The purpose of this chapter is to present the methods in a way that makes it possible to replicate the study and the results. The methods are partially complemented by examples in the theory section. Chapter 5 presents the results of the study that are relevant to answer the initial problem statement. This includes various measurements of predicted financial sentiments and stock returns and the relationship between the variables. The results are discussed based on the research questions and the theoretical framework in Chapter 6. This chapter also intends to place the study in the existing research and discuss the contributions to the research field. Finally, a conclusion is given in Chapter 7, which summarizes the main results and provides suggestions for improvement and future research.

2 Theory

This chapter provides an overview of the relevant underlying theory of the methods and the discussion of the thesis. Relevant concepts of the stock market, such as industry standard, asset strategies, and financial reporting are explained. Additionally, the theory behind each step of the data pipeline will be explained in detail to understand the underlying mechanism of text extraction, sentiment analysis, and visualization. Furthermore, the measurements applied in the study are explained in this chapter. In addition, the details of the applied Python libraries are explained along with relevant code examples.

2.1 The Stock Market

To understand the scope of this thesis, a brief introduction to the stock market is needed. The stock market is a broad term and typically means the gathering of exchanges and venues where shares of companies can be bought and sold. The Oslo Stock Exchange (OSE) is an example of a stock exchange being a part of the overall stock market. The way the stock market is structured is important to maintain a free market economy so that anyone can invest and exchange capital. Due to the rise of electronic communication, most trades in stocks and financial instruments occur almost instantly [26].

2.1.1 Global Industry Classification Standard

The stock market, and more specifically the companies listed on a stock exchange, can be grouped into different industries and sectors through the Global Industry Classification Standard (GICS). The Oslo Stock Exchange follows the standard and is divided into eleven sectors; Technology, Telecommunications, Health Care, Financials, Real Estate, Consumer Discretionary, Consumer Staples, Industrials, Basic Materials, Energy, and Utilities [27]. The different industries are divided into sectors, e.g. one sector within the energy industry is oil, gas, and coal.

2.1.2 Asset Strategy

A short definition of “asset-light” and “asset-heavy” is necessary to get an understanding of the choice of companies and further the discussion on it. Generally, asset-heavy companies would generate income by utilizing their fixed assets. Asset-light companies, on the other hand, own fewer capital assets and typically have advantages compared to asset-heavy companies, as they can be more agile and reduce the risk. There may exist both asset-heavy and asset-light companies within the same industry because of different business models. Within real estate, Airbnb is a good example of an asset-light company, and it is also close to being an asset-zero structure. Companies, such as Opendoor, are buying the properties and are therefore considered asset-heavy.

2.1.3 Quarterly Report Standard

For listed companies, it is mandatory to publish quarterly reports to inform and safeguard the interest of investors. A quarterly report gives information about the performance of a company and will typically include income statements and balance sheets. In addition, an executive summary, highlights, and future objectives are typically included. The Oslo Stock Exchange follows the IFRS standard IAS 34 for mandatory reporting [28]. The mandatory parts are statements of financial position, comprehensive income, changes in equity, cash flows, and selected explanatory notes [29]. These interim reports are obligated to report the mentioned numbers and compare them with the results of the last period.

2.2 Data Pipeline

The popularity of using data in product development is increasing [30]. Data products like dashboards and benchmarks are of importance in developing machine learning models and in decision-making processes. Data pipelines work well for detecting errors and for the efficient development of data products.

A data pipeline is a chain of processing elements in an order in which the output of one element can be the input of the next element. This provides an opportunity to handle and transform data with different operations so that the desired output format can be generated. Data pipelines are implemented in such a way that the overall efficiency of data flow can be increased. More specifically, the flow from the source to the destination reduces the need for humans in the loop.

The starting point of a data pipeline is often a source that generates data. The final module of the pipeline receives the processed data. An optimal ending point deploys a visualization tool or machine learning model. Data pipelines can be constructed to handle different data types and size of data batches. Abstract representations of end-to-end data pipelines are often referred to as ETL pipelines, consisting of the three activities; Extract, Transform and Load data. The particular pipelines built for this thesis are illustrated and described in Chapter 4.

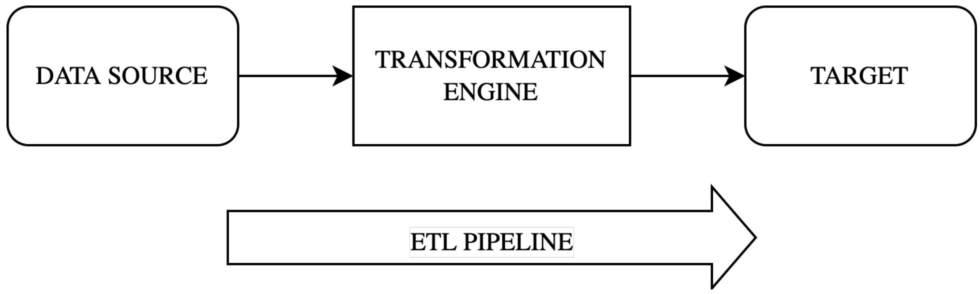


Figure 2.1: Illustration of elements in an ETL pipeline. Data is extracted from a data source, transformed in a transformation engine and loaded to a target.

2.3 Text Extraction

Data extraction is essential for a data pipeline. Data can be extracted through web scraping, APIs, cloud storage, or locally stored files. Additionally, specific data types must be extracted from the main data source. In this study, the text extraction from PDF files is an essential process in the data pipeline. To extract text from PDF files, the Python package PDFMiner is applied.

2.3.1 Portable Document Format

Text data can be stored in different file formats. One of the most common formats is Portable Document Format (PDF). PDF can be considered a digital paper and is used for easy distribution and readability. It represents the visual appearance of electronic documents in applications, software, hardware, and operating systems [31]. PDFs are compressed, which means that some of the original data types and sources are lost when creating a PDF. The compression is favorable for distribution and storage, but involves challenges for precise text extraction because the text is not stored in a plain text format, but rather as a layout-based format [32]. In a PDF the positions and fonts of individual characters are specified, but identifying words is non-trivial because of numerous reasons.

The variation in spacing, word order, paragraph boundaries, and semantic roles constitutes the challenge of extracting and cleaning text.

2.3.2 PDFMiner

PDFMiner is a Python package that is used to extract information from PDF documents. This package parses all objects from a PDF document into different Python objects. The texts are grouped for readability. In addition to the extraction of text, images, tags, and tables can be extracted using PDFMiner. This tool extracts the text directly from the PDF's sourcecode. In addition to this, it can be used to extract the font of the text [33]. In this thesis, a community-maintained fork of the PDFMiner, pdfminer.six, is applied for text extraction. It is a tool for extracting elements from PDF documents and can be used to extract font sizes from text. New features are actively added to improve usability. The example below shows the extraction of an element in a PDF using pdfminer.six.

Listing 1: Example of element extraction using pdfminer.six

```
from pdfminer.high_level import extract_pages
for page_layout in extract_pages("EQNR_Q1_2019.pdf"):
    for element in page_layout:
        print(element)
```

Pdfminer.six focuses on the extraction and analysis of text. In Python, data can be extracted directly from PDF files. The package offers functionality for extraction of tables, forms, and illustrations. Pdfminer.six applies an analysis algorithm for layouts (LA) [34]. This algorithm can be divided into three steps; character grouping, line arranging, and box ordering. The text is returned by the hierarchically arranged boxes. The PDFParser method is used to parse the PDF file as it is loaded as a binary object.

2.3.3 Compiling with Regular Expressions

Regular expressions are specifications of a set of strings for matching. That means that regular expression functions search for given regular expressions to match strings. In Python, the re module offers a regular expression engine, which can be applied to compile regular expressions into objects and perform matches. Furthermore, this is in particular useful for various operations, such as searching for the start and end of sentences by applying the built-in findall() method. This method returns all substrings where the regular expression matches and returns them as a list. An example of using regular expressions in Python is presented in Listing 2.

Listing 2: Example of regular expression for string matching in Python

```
regex_compiler = re.compile(r'([A-Z][^\.!?]*[\.!?])', re.M)
extracted_sentences = regex_compiler.findall(raw_text)
```

2.4 Data Transformation

Extracting text can be seen as both an extraction process and data transformation as it changes the format, structure and value of the data. This study defines transformation as the process of transforming extracted and cleaned text into predicted sentiment scores. The transformation is necessary to increase the efficiency of the analytical process and is often crucial to enable data-driven decision-making. Machine learning and deep learning are often applied in this process to transform data with the use of algorithms.

2.4.1 Machine Learning and Deep Learning

decision-making

2.4.2 Natural Language Processing

Natural Language Processing (NLP) is a part of AI and focuses on providing computers with the ability to understand language in a natural way. NLP can use both ML and DL approaches to execute language modeling. Common tasks include recognizing speech, finding the sentiment in a text and generating responses to questions. Different deep learning architectures are applied for NLP tasks. One of these is the bidirectional long short-term memory based on a recurrent neural network [35]. Transformers is another architecture applied for language models like BERT. The transformers architecture enables modeling of text in bidirectional manner. That means that the network learns context and relationships in sequential data [36].

2.4.3 Tokenization

Tokenization is an important step in most NLP tasks. It is a process of splitting text documents into sentences, terms, words, or symbols. The tokenization results in numerous tokens, which are usually separated by a full stop or a white space. Furthermore, tokenization is a method of vectorizing words into a sequence of numbers, which is a format supported by ML models [37]. The tokenizer used in this thesis is a BertTokenizer which is based on WordPiece and PreTrainedTokenizer.

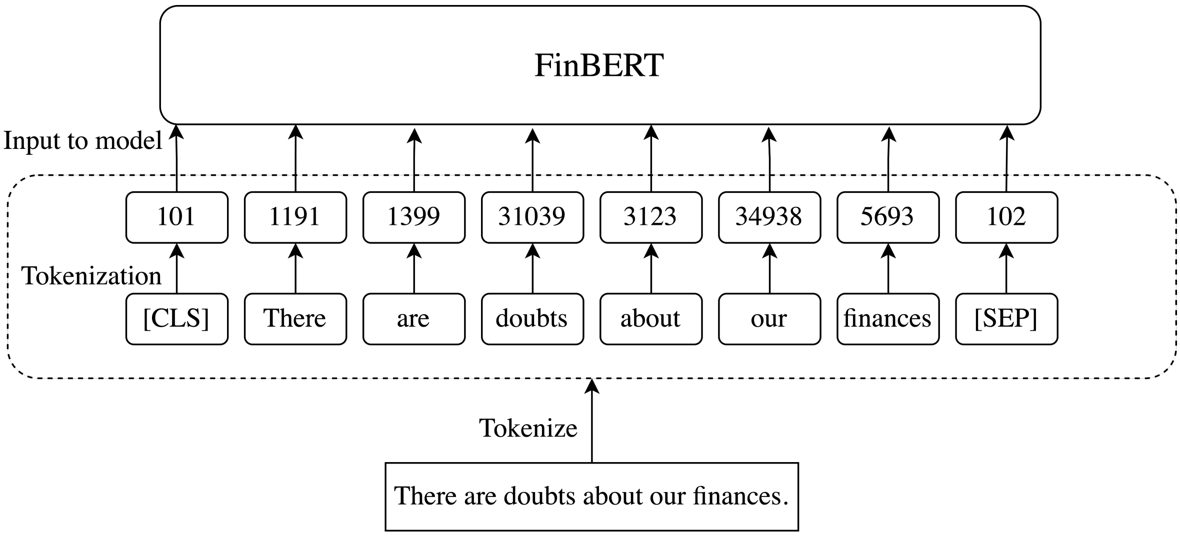


Figure 2.2: Illustration of the BERT tokenization

The example above, Figure 2.2, illustrates the tokenization with the BertTokenizer. The model classifies sentences, and therefore special tokens are needed for the starting and ending positions of a sentence. These tokens are named [CLS] and [SEP].

2.4.4 Sentiment Analysis

Sentiment analysis is the process of identifying and categorizing opinions from text using computers [6]. This can make it possible to uncover the author's attitude toward an entity. Other words similar to sentiment are used interchangeably in the literature, such as affect, feeling, emotion, and opinion [25]. Sentiment analysis is a method that makes it possible to analyze and judge text to extract the opinion. One should be careful with opinion mining, as it is a difficult task to extract insight into interconnections between emotions. The automation of sentiment analysis makes it possible to predict the polarity based on models trained for a certain domain.

The sentiment can be categorized as positive, negative, and neutral. This information can be used for insight and decision-making. A sentiment classification can be applied to financial text and is often referred to as financial tone analysis [38]. The development of financial language models with DL architecture enables contextual understanding of whole sentences.

2.4.5 Pre-trained Language Models

A recent discovery in the research of language modeling is that fine-tuning existing language models can be successful for many NLP tasks [15]. The advantage of pre-trained language models is that the base of the models is built on large collections of text data, also called corpora. By adding a new layer to the model, the model can be fine-tuned for a specific domain.

2.4.6 Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a deep learning architecture consisting of a stacked encoder layer from the transformer. The idea behind the BERT model is an innovative way to classify text. In simple terms, the BERT model is trained by masking words in a sequence and by classifying sentences that follow each other. The model consists of two steps, pre-training and fine-tuning [35]. During pre-training the model is trained with unlabeled data. Fine-tuning is done using labeled data from downstream tasks. These downstream tasks can be such as Question Answering (QA) and Natural Language Inference (NLI). These tasks are based on understanding the relationship between sentences. BERT is a language model designed to pre-train deep bidirectional representations from unlabeled text. The model is designed to pre-train deep bidirectional representations by conditioning unlabeled text on both the left and right contexts of the layers. BERT is in concept simple, but efficient and powerful. It gives state-of-the-art results on several NLP tasks. Furthermore, the model can be fine-tuned with an additional layer for a specific task. The BERT model architecture is a multi-layer bidirectional transformer encoder [7]. The prediction task of training the BERT model is illustrated below.

Figure 2.3 shows how the word "doubts" is predicted. First, the words are transformed

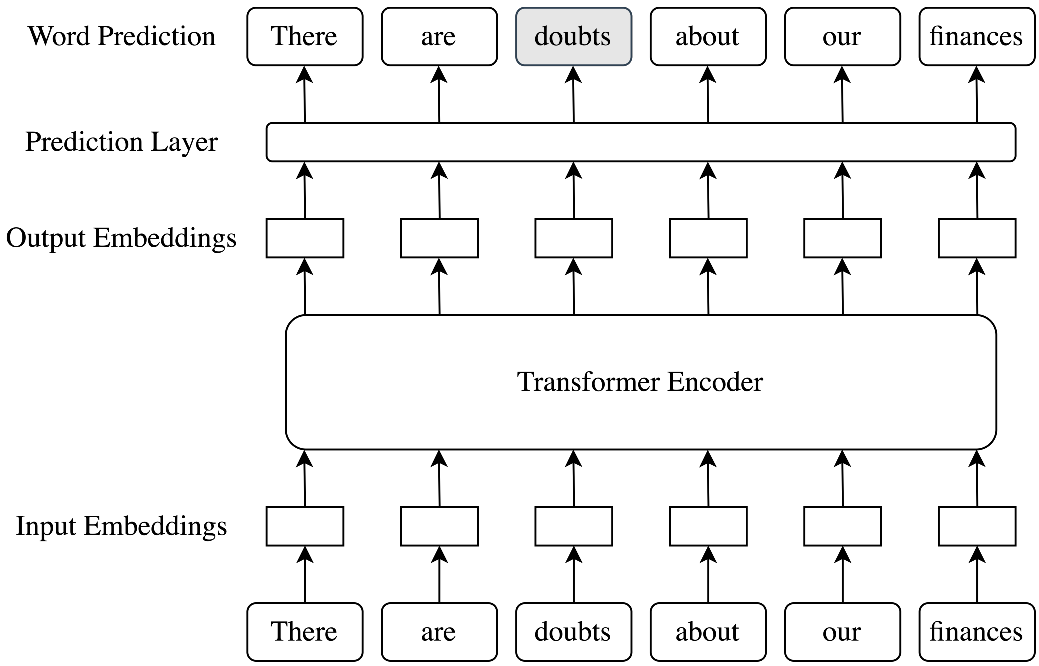


Figure 2.3: Illustration of the BERT prediction task

into a vector representation in the embedding and transformed with the transformer encoder. The generated output is the words represented by new vector values in the huge space of word vectors. This enables the prediction of the masked words and the fine-tuning of additional layers. An example is the fine-tuning of a financial classification layer to develop the domain-specific model, FinBERT.

Traditionally, NLP models receive a single word token as input, but the BERT model learns the meaning hidden between words, as it handles whole sentences as input. In this way, the sentiment can be predicted for an entire sentence in a given context.

2.4.7 FinBERT

The previous section presented the underlying principles of the BERT model. FinBERT is a domain-specific BERT model developed at the Hong Kong University of Science and Technology. The model is pre-trained on large financial corpora of 4.9 billion tokens. These include corporate reports, earnings conference call transcripts, and analyst reports. According to experiments by the developers of the model, FinBERT outperforms the more generic BERT-models for financial sentiment classification [39]. Furthermore, the model is a good choice for financial sentiment analysis given the demonstrated performance in the study. FinBERT predicts the sentiment of a given tokenized sentence.

According to the software documentation [39], FinBERT is state-of-the-art when it comes to performance for financial sentiment tasks. The model applied in this thesis is a fine-tuned FinBERT model, which is fine-tuned on 10,000 analyst statements for tone prediction. The following example from the official GitHub documentation [40] shows how tokenization and sentiment prediction can be performed using FinBERT and Python.

Listing 3: Predicting financial sentiment with FinBERT. Sentences are tokenized and predicted as negative, positive or neutral.

```
from transformers import BertTokenizer,
                              BertForSequenceClassification
import numpy as np

finbert = BertForSequenceClassification.from_pretrained(
            'yiyanghkust/finbert-tone',num_labels=3)
tokenizer = BertTokenizer.from_pretrained('yiyanghkust/finbert-tone')

sentences = ["there is a shortage of capital, and we need extra financing",
             "growth is strong and we have plenty of liquidity",
             "there are doubts about our finances",
             "profits are flat"]

inputs = tokenizer(sentences, return_tensors="pt", padding=True)
outputs = finbert(**inputs)[0]

labels = {0:'neutral', 1:'positive',2:'negative'}
for idx, sent in enumerate(sentences):
    print(sent, '----', labels[np.argmax(
    outputs.detach().numpy()[idx])])

'''
there is a shortage of capital, and we need extra financing ---- negative
growth is strong and we have plenty of liquidity ---- positive
there are doubts about our finances ---- negative
profits are flat ---- neutral
'''
```

The FinBERT model uses cross-entropy as the loss function, and the pre-training of the model takes about two days with one million iterations until the loss converges.

The intention of pre-training models is that financial practitioners and researchers can benefit from the model without the need of significant computational power, such as the application of FinBERT in this study.

2.4.8 Softmax

For classification, FinBERT uses a simple linear layer with a softmax activation function. The softmax function scales the predicted sentiment scores into probabilities that sum up to 2. The function can be seen as a generalization of logistic regression for multi-class classification. The softmax function is useful because it normalizes the values in the final output layer of a neural network. Additionally, softmax activation for the FinBERT model enables the return of the sentiment label on the basis of the classification scores of the input token. Mathematically, the softmax formula is presented below.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K$$

Where,

$$\sigma = softmax\ function$$

$$z_i = value\ of\ neuron$$

$$e^{z_i} = exponential\ function$$

$$\sum_{j=1}^{K} e^{z_j} = normalization\ term$$

$$K = number\ of\ classes\ in\ the\ classifier$$

In Python, the NumPy argmax function can be used to return the indices from the element with the maximum value in an array. By mapping the indices to a dictionary, the label of a classified sentence can be assigned. The following Listing 4 shows the values from the FinBERT classification and the label assigned to the argmax function, which is demonstrated in the Listing 3.

Listing 4: Assigning label from softmax ouput. The negative label is assigned as the weight for a negative label is the largest.

```
labels = {0:'neutral', 1:'positive',2:'negative'}
there are doubts about our finances [-1.9989771 -4.5428824 8.526565 ]
there are doubts about our finances ——— negative
```

2.4.9 Stock Performance

To measure stock performance, a simple and efficient calculation of stock return is used in this thesis. A simple and efficient way of measuring stock performance is to find the ending price subtracted by the starting price and divide this number by the starting price. This can be expressed as follows.

$$r_j = \frac{P_{t+1} - P_t}{P_t}$$

Where,

$$r_j = Return\ of\ stock\ j$$

$$P_t = Price\ of\ the\ stock\ at\ time\ t$$

$$P_{t+1} = Price\ of\ the\ stock\ at\ time\ t + 1$$

The method is simple, but easy to implement as it requires nothing more than the price of a stock for specific time periods. In financial data, stock prices are often divided into four categories of prices: bid, ask, open and close. Bidding and asking prices are the highest bid price and the lowest ask price for the stock on a given day. The opening price is the price for the first trade of the day, and the closing price is the last traded price.

To calculate the one-day return of a stock, one can simply point out the closing price of one day as the starting price and the closing price of the next day as the ending price. Furthermore, for one-week returns, the same method can be applied, but by using the closing price for the stock seven days after the starting price. A one-quarter return would typically mean a return on the stock after three months. An approach to calculate this number is to use the opening price of the first active trading day of the starting month of the quarter as the starting price, and the closing price of the last active day of the third month as the ending price. This data can be downloaded through the Yahoo! finance API with the yfinance tool in Python.

2.4.10 yfinance

Yahoo! Finance is a network which provides financial news and data. The network has a publicly available API which can be used through the open-source tool yfinance. The Python tool enables downloading of market data from Yahoo! finance in a fast and reliable way through different modules. Downloading market data can be done using the download module presented by the example in Listing 5. It returns data such as date, opening price and closing price for a given stock. Additionally, the tool can be used to download a wide range of other stock-specific data.

Listing 5: Downloading data for Equinor with yfinance for specified start and ending date.

```
import yfinance as yf
data = yf.download("EQNR.OL",start="2019-01-01",
                              end="2021-12-31")
```

2.4.11 Normalization

Normalization is a statistical scaling method. The goal is to transform values into a similar scale. The method is for instance essential for the visualization of features with different scales. Additionally, normalization is important for machine learning tasks as most algorithms operate under the assumption of training and test data being drawn from the same distribution [41].

In this thesis, the data are scaled with Mean Normalization. In simple terms, Mean Normalization is a method to normalize the range of independent variables of data by calculating and subtracting the mean for every variable. This value is either divided by the range maximum-minimum range value or the standard deviation. In this thesis the standard deviation is used, which is commonly named standardization. The formula for the standardization is presented below.

$$x' = \frac{x - \mu}{\sigma}$$

Where,

$$x' = standardized\ variable$$

$$x = variable$$

$$\mu = mean\ of\ x$$

$$\sigma = standard\ deviation\ of\ x$$

In Python, this method can be used on DataFrames using the Pandas built in functions for mean and standard deviation, as presented in Listing 6 where x is tabular data. Pandas is a Python package providing tools for data analysis and manipulation in a fast and intuitive way [42]. Pandas DataFrame is a 2-dimensional data structure similar to a table with rows and columns.

Listing 6: Normalizing Pandas DataFrame columns in Python with the standardization method

```
import pandas as pd

values = [[1, 5, -10], [2, 8, -20], [3, 15, -40]]

x = pd.DataFrame(values)
x_normalized=(x-x.mean())/x.std()
```

This would result in a scaling of numbers as illustrated in Table 2.1.

	0	1	2
0	1	5	-10
1	2	8	-20
2	3	15	-40

$\xrightarrow{Normalization}$

	0	1	2
0	-1.0	-0.84	0.87
1	0.0	-0.26	0.22
2	1.0	1.10	-1.09

Table 2.1: Normalizing values

2.5 Validation

Validation is necessary to measure the precision of the methods and their intentions. High validity refers to methods that correspond to real values, characteristics, and variations. A common method for validation of machine learning models and sentiment prediction is the accuracy. Additionally, correlation measures the relationship between variables and is used to assess relative validity.

2.5.1 Correlation

As a measurement of the relationship between two different variables, Pearson correlation coefficient (r) is the most common way to quantify the relationship [43]. It is a measure of linear association, which has a value from -1 as a perfect negative linear correlation to 1 as a perfectly positive linear correlation. A more positive correlation coefficient is interpreted as greater validity. The strength of the relationship can be divided into four levels for the absolute correlation value, as presented in Table 2.2.

Absolute value of r	**Strength of relationship**
r <0.25	No relationship
0.25 <r <0.5	Weak relationship
0.5 <r <0.75	Moderate relationship
r >0.75	Strong relationship

Table 2.2: Coefficient values for the strength of the relationship

The formula for the correlation coefficient is defined by the following formula.

$$r = \frac{Cov(x,\ y)}{\sigma_x\ \sigma_y}$$

Where,

$$r = Pearson\ correlation\ coefficient$$

$$Cov\,(x,\ y) = covariance\ of\ variables\ x\ and\ y$$

$$\sigma_x = standard\ deviation\ of\ x$$

$$\sigma_y = standard\ deviation\ of\ y$$

In this study, the Pandas DataFrame Pearson correlation method, corr(), has been used, which computes the pairwise correlation of columns.

2.5.2 Accuracy

Accuracy is a common metric for machine learning applications. Generally, it is the number of correctly predicted data points out of all data points in the prediction. To measure the accuracy of sentiment prediction in a case where the neutral labels are less interesting, a common method is to evaluate the positive and negative labels [23, 37]. Considering the classified labels as true positive (TP), true negative (TN), false positive (FP) and false negative (FN), the accuracy is defined as presented below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2.5.3 Misclassification

Misclassification is another measurement which is defined as the sum of incorrect predictions divided by the total number of predictions. In simple terms, it gives the percentage of incorrectly predicted samples. The measurement is used to evaluate models where reducing error is essential, such as fraud detection [44].

$$Misclassification = \frac{FP + FN}{TP + TN + FP + FN}$$

2.5.4 Precision

Precision calculates the effectiveness of a classification task and is the number of TP divided by the sum of TP and FP. It therefore measures the ability of the model to classify positive samples, and is often in tension with recall as high precision typically means low recall [45].

$$Precision = \frac{TP}{TP + FP}$$

2.5.5 Recall

Recall is another measurement for effectiveness which gives the fraction of correctly predicted positives for the total number of TP and FN. While precision is a measurement of quality, recall can be considered a measurement for quantity.

$$Recall = \frac{TP}{TP + FN}$$

2.5.6 F1-Score

The F1-score gives another measurement of incorrectly classified labels. It is a measurement suitable for inbalanced class distribution. It is calculate as presented in the formula below. The measurement is widely used to evaluate models as it sums up the predictive performance by evaluating both precision and recall [46].

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

3 Data and Materials

This chapter gives an overview of the data sources used in this thesis and the reasons for the chosen data. Further descriptions of the data are presented with the methods in Chapter 4. The data are mainly collected from the investor relations page of the chosen companies and through the Python tool yfinance.

3.1 Quarterly Reports

The quarterly reports are collected from real estate companies listed on the Oslo Stock Exchange (OSE) from their respective websites. The reports downloaded range from the first quarter of 2019 to the fourth quarter of 2021.

The reports analyzed in this thesis are the quarterly reports of companies which are all listed at the OSE and the reports are written in English. The companies can be divided into asset-heavy and asset-light companies. These companies are large-cap companies across different industries. In this thesis, technology, telecommunications, real estate, and oil/gas are the chosen industries. The "asset-heavy" companies are within real estate and oil/gas. The asset-light companies operate in the technology and telecommunications industry.

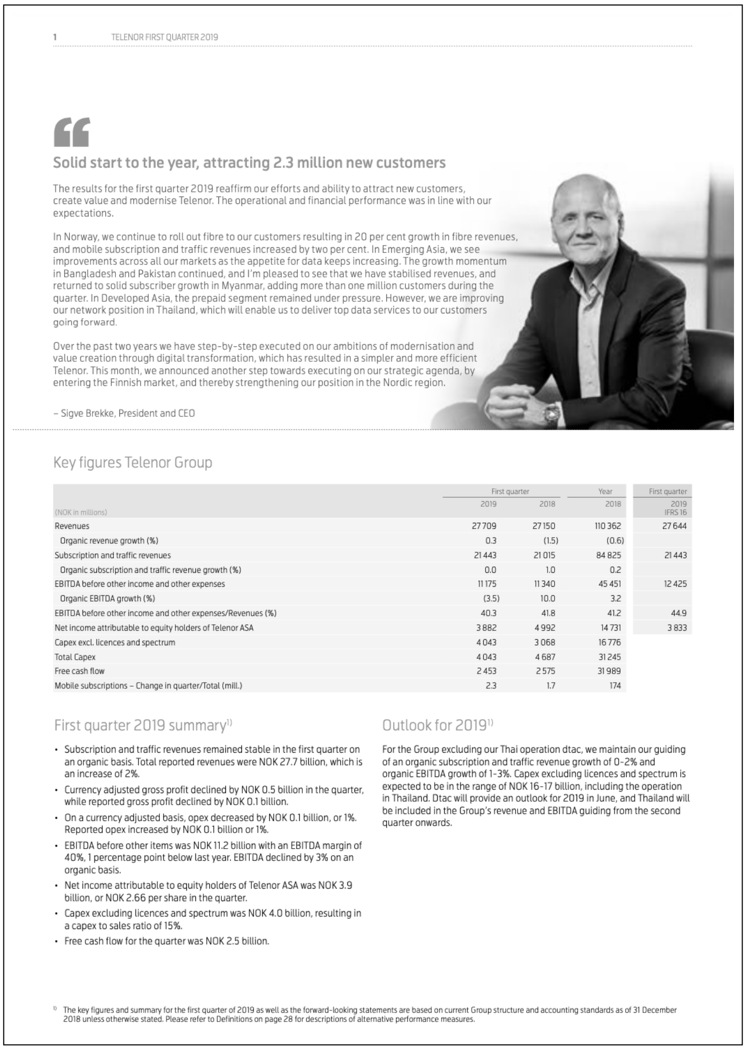
As described in Chapter 1, previous research claims that stock forecasting for asset-heavy companies is more promising than forecasting for asset-light companies [22]. Therefore, two asset-heavy industries and one asset-light industry have been chosen for comparison. The companies are large companies and are chosen because of accessible data, such as the quarterly reports themselves, but also news and comments related to the company, which makes it easier to interpret results. The selected companies are presented in Table 3.1. For visual reasons, Technology/Telecommunications will be referred to as Technology throughout the thesis.

Furthermore, quarterly reports are collected from 2019 to 2021. For each year, four reports are collected (Q1-Q4), which in total gives 12 interim reports for each company. The only exception is Kahoot!, which was not listed on OSE before late 2019, and therefore only has reports from the third quarter of 2019 and beyond. If accessible, the dates for when

Technology/Telecommunications	**Real Estate**	**Oil/Gas**
Asset-light	*Asset-heavy*	*Asset-heavy*
KAHOOT! ASA	ENTRA ASA	EQUINOR ASA
SCHIBSTED ASA	OLAV THON EIENDOMSSELSKAP ASA	NEL ASA
TELENOR ASA	SELVAAG BOLIG ASA	AKER BP ASA

Table 3.1: The selected companies and industries from OSE grouped by industry and asset strategy.

the reports were published have also been collected. For companies without publication dates, the dates have been collected from the message history on the Euronext webpage. In general, the reports are of length 30-40 pages. The layout of the reports varies for different companies and quarters, but typical for all of them is the inclusion of tables and illustrations. A typical page layout of a quarterly report of Telenor ASA is provided in Figure 3.1.



1 TELENOR FIRST QUARTER 2019

Solid start to the year, attracting 2.3 million new customers

The results for the first quarter 2019 reaffirm our efforts and ability to attract new customers, create value and modernise Telenor. The operational and financial performance was in line with our expectations.

In Norway, we continue to roll out fibre to our customers resulting in 20 per cent growth in fibre revenues, and mobile subscription and traffic revenues increased by two per cent. In Emerging Asia, we see improvements across all our markets as the appetite for data keeps increasing. The growth momentum in Bangladesh and Pakistan continued, and I'm pleased to see that we have stabilised revenues, and returned to solid subscriber growth in Myanmar, adding more than one million customers during the quarter. In Developed Asia, the prepaid segment remained under pressure. However, we are improving our network position in Thailand, which will enable us to deliver top data services to our customers going forward.

Over the past two years we have step-by-step executed on our ambitions of modernisation and value creation through digital transformation, which has resulted in a simpler and more efficient Telenor. This month, we announced another step towards executing on our strategic agenda, by entering the Finnish market, and thereby strengthening our position in the Nordic region.

– Sigve Brekke, President and CEO

Key figures Telenor Group

	First quarter		Year	First quarter
(NOK in millions)	2019	2018	2018	2019 IFRS 16
Revenues	27 709	27 150	110 362	27 644
Organic revenue growth (%)	0.3	(1.5)	(0.6)	
Subscription and traffic revenues	21 443	21 015	84 825	21 443
Organic subscription and traffic revenue growth (%)	0.0	1.0	0.2	
EBITDA before other income and other expenses	11 175	11 340	45 451	12 425
Organic EBITDA growth (%)	(3.5)	10.0	3.2	
EBITDA before other income and other expenses/Revenues (%)	40.3	41.8	41.2	44.9
Net income attributable to equity holders of Telenor ASA	3 882	4 992	14 731	3 833
Capex excl. licences and spectrum	4 043	3 068	16 776	
Total Capex	4 043	4 687	31 245	
Free cash flow	2 453	2 575	31 989	
Mobile subscriptions – Change in quarter/Total (mill.)	2.3	1.7	174	

First quarter 2019 summary[1)]

- Subscription and traffic revenues remained stable in the first quarter on an organic basis. Total reported revenues were NOK 27.7 billion, which is an increase of 2%.
- Currency adjusted gross profit declined by NOK 0.5 billion in the quarter, while reported gross profit declined by NOK 0.1 billion.
- On a currency adjusted basis, opex decreased by NOK 0.1 billion, or 1%. Reported opex increased by NOK 0.1 billion or 1%.
- EBITDA before other items was NOK 11.2 billion with an EBITDA margin of 40%, 1 percentage point below last year. EBITDA declined by 3% on an organic basis.
- Net income attributable to equity holders of Telenor ASA was NOK 3.9 billion, or NOK 2.66 per share in the quarter.
- Capex excluding licences and spectrum was NOK 4.0 billion, resulting in a capex to sales ratio of 15%.
- Free cash flow for the quarter was NOK 2.5 billion.

Outlook for 2019[1)]

For the Group excluding our Thai operation dtac, we maintain our guiding of an organic subscription and traffic revenue growth of 0-2% and organic EBITDA growth of 1-3%. Capex excluding licences and spectrum is expected to be in the range of NOK 16-17 billion, including the operation in Thailand. Dtac will provide an outlook for 2019 in June, and Thailand will be included in the Group's revenue and EBITDA guiding from the second quarter onwards.

1) The key figures and summary for the first quarter of 2019 as well as the forward-looking statements are based on current Group structure and accounting standards as of 31 December 2018 unless otherwise stated. Please refer to Definitions on page 28 for descriptions of alternative performance measures.

Figure 3.1: A layout example from a page in a quarterly report of Telenor ASA.

3.2 Financial Data

Stock data is collected with the Python tool yfinance as described in Chapter 2 and used for stock performance computation for different time intervals before and after the date of publication of the report. yfinance is chosen as it is a fast and reliable tool for downloading financial data from a well established financial data provider, Yahoo! finance.

3.3 Possible Sources of Errors

The dates that are read and collected manually may be a source of error. Despite careful work and several reviews, there could still be sources of errors, such as key-entry mistakes. Additionally, the filenames of the downloaded reports do not have the same standard. This makes processing and storing the results challenging. To aid in the task, every file has been renamed to the format *company_quarter_year*. However, this may be another source of error as it can result in key-entry errors. A further discussion of possible errors and data quality is presented in Chapter 6.

4 Methods

This chapter presents the methods applied in order to answer the research questions. Furthermore, this chapter gives a detailed overview of how the data is flowing through the data pipeline and discusses each step in such a way that it should be possible to execute the same process using Python. The methods presented in this chapter are partly complemented by code examples in the theoretical framework presented in Chapter 2.

As the methods build on an ETL pipeline structure, the methods can briefly be separated into four categories; Extraction, Transformation, Loading, and Deployment. Data extraction is done by loading locally stored PDF files which are manually downloaded quarterly reports. The transformation is done with different methods such as PDF mining, tokenization, and financial sentiment analysis. The results are loaded into locally stored pickle files. The deployment is done through visualization of results descending from the output files.

The methodology for this thesis work can be divided into two pipelines. The first constructed pipeline reads PDF reports and extracts the sentiment analysis. The reason for splitting the data flow into two pipelines is that the sentiment pipeline can be used for any financial report to derive financial sentiment from any PDF file. The second pipeline represents the other part of the thesis and the research question considering how the extracted sentiment can be compared to the market and stock movements. This pipeline is a visualization pipeline that downloads relevant stock data based on the publication date for quarterly reports.

4.1 The Sentiment Pipeline

The sentiment pipeline takes in the raw PDF reports, which are locally stored files. Furthermore, the text is extracted using a Python library named PDFMiner. The next step of the pipeline is to determine the financial sentiment of each sentence using the FinBERT model. As a final step, the sentiment scores are stored as pickle files. These files are the input for the second pipeline, explained in the next subchapter. Figure 4.1 gives an overview of the elements of the financial sentiment pipeline.

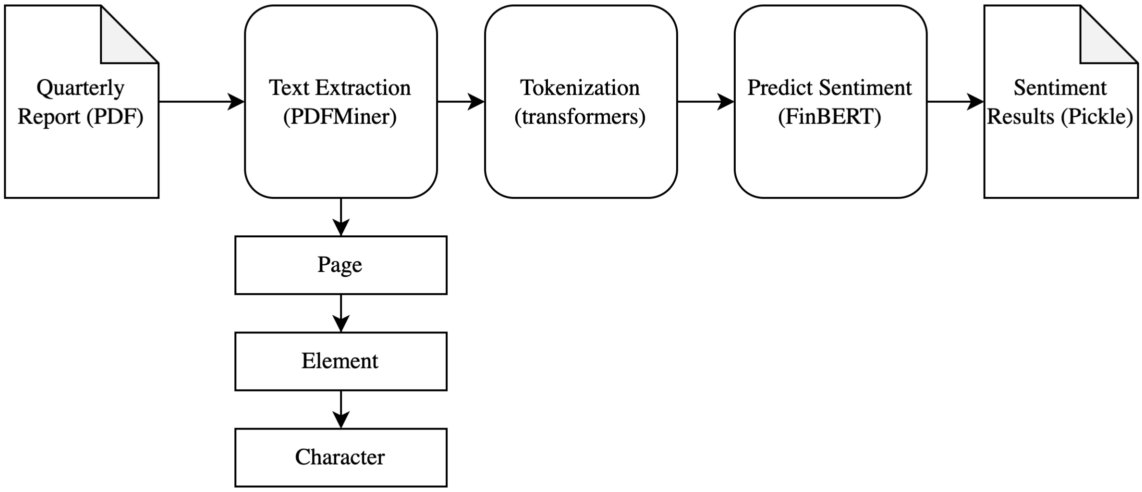


Figure 4.1: Illustration of pipeline for text extraction and sentiment prediction. Quarterly reports are collected from the internet as PDF-files. The text from the report is extracted and cleaned by iterating characters. Tokenization vectorize the cleaned text so that sentiment prediction can be performed with FinBERT. The results are loaded as pickle files locally.

4.1.1 Downloading Data

As elaborated in the previous chapter, quarterly reports have been manually downloaded from the investor's page from the company websites as PDF files. These have been organized into folders for company, year, and quarters. The process could have been carried out automatically by constructing a pipeline with web scraping and automatic download from the desired companies. However, it has not been focused on in this thesis because of limited time and relevance to the research questions.

4.1.2 Extracting Text from PDF-documents

The reading of the data is performed by uploading the local files with Python. Further, the pipeline applies PDFMiner to extract the elements from the reports, such as figures, pictures, and text. The text data are extracted with the tool by extracting every recognized text element. Furthermore, the text data is analyzed into characters, whereas the characters can be divided into font sizes and font types. This helps to identify the size and font of the letter used in the common text of the report. It enables the exclusion of irrelevant characters, which can be considered noise. This is a challenging task because of the flexible PDF layout. Therefore, regular expressions need to be constructed to extract and clean the text.

This step can be considered as both an extraction step and a transformation step. The data format (PDF) is transformed into Python text strings. However, considering this operation as text extraction, it is reasonable that it belongs to the extraction process of the pipeline.

The extracted text is transformed into sentences based on regular expression functions, as described in Chapter 2.3.3. These sentences are the input sentences for tokenization.

4.1.3 Tokenization

Tokenizing the words of the extracted text is needed to break down the text into smaller chunks, tokens. These tokens typically represent a word and help analyze the sequence of words and are therefore essential for predicting the sentiment. There are several Python tools for tokenization of text, but in this case the BertTokenizer from the transformers library has been used. The reason for this is the compatibility with BERT and FinBERT. The tool is optimized for these models by tokenizing whole sentences [47].

4.1.4 Predicting the Sentiment

After the characters have been extracted and stored in a list, the characters are joined as long strings that can be read by a language model. This results in a long string in Python which then can be tokenized. The FinBERT model is then used in the Python pipeline to extract the sentiment of each sentence of the extracted and cleaned text from the report. This results in a sentiment score for each sentence. With an argmax-function, it is possible to turn the numbers into positive, negative, or neutral labels. The final step is to use a Python Counter function to find the number of neutral, negative, and positive labels assigned to the report.

4.1.5 Saving Results

Each file is processed and transformed with the pipeline, and the output for each processed report is a score summary containing the number of neutral, negative, and positive labels. It also contains the average neutral, positive, and negative scores for the entire report. The last output format is the predicted sentences with the predicted labels needed for the accuracy evaluation.

To process several files at one time, this pipeline iterates all files stored for a single company in a local file folder. In this way a result file can be generated which contains the results of all quarterly reports with their respective scores.

4.2 The Comparison Pipeline

The second pipeline is needed for comparing the predicted financial sentiment scores for each report with the stock performance of a company. This is needed to answer the research questions regarding the actual performance of the company for the period, but also the one-day and one-week return from the publication date. This pipeline takes in publication dates for each quarter for a chosen company and merges them into the stored results. The data flows through the pipeline and sentiment scores are transformed into a single score for each report. Financial data is downloaded with the publish date or beginning of the quarter as a starting point. The processes are explained in detail throughout this section. Figure 4.2 gives an overview of the visualization pipeline.

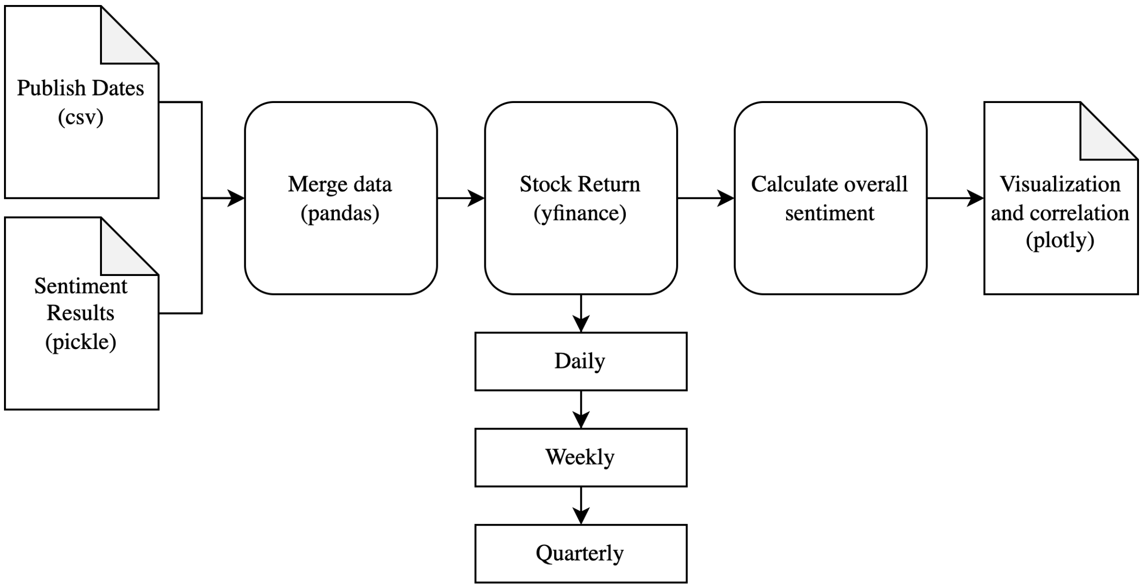


Figure 4.2: Illustration of pipeline for comparing stock performance and sentiment scores. Collected publish dates and sentiment results are merged with pandas. The dates are used for downloading and calculating stock returns. Sentiments for each report is computed and compared to the stock performance visually.

4.2.1 Input Files

The first input file for this pipeline is a file that contains the dates for when the reports were published for each quarter. This file also contains the start date of a quarter if the one-quarter return shall be calculated. The second file is the output file from the previous pipeline which predicts the sentiment scores for each sentence extracted from the report. In addition, the file contains a column that describes the specific quarter and year from which the report is derived.

4.2.2 Merging Data

The next process in the pipeline is the merging of the two files on the quarter column. This gives one dataset in the pipeline containing the results and date for each quarterly report. These data are sent to the next step of the pipeline to calculate the return of the stock.

4.2.3 Calculating Stock Return

The stock return is calculated in different ways depending on specific input arguments. One can receive an output of one-day, one-week, or past one-quarter stock return. For all these variants, the data is downloaded with yfinance. The general calculation of the stock return is in simple terms a subtraction of the end price from the starting price, divided by the starting price as explained in Chapter 2.

The one-day return is calculated by taking the closing price of the day before the report is published as the starting price and the closing price of the day the report is published as the ending price. The one-week return takes the close price the day before the report is published as the starting price and the close price 7 days later as the ending price. The last calculation is a one-quarter return from the past quarter, which means the quarter that is being reported. This is done by finding the first open market day backward from the last day of the quarter. The starting date is the first open market date in the first month of the quarter, and the opening price of the day is the starting price. These are all simple methods for stock market returns, but may be sufficient for the scope of this thesis.

4.2.4 Calculating Overall Sentiment

The next step of the pipeline is the calculation of a sentiment score from the labeled sentences of the report. This is done by assigning values to the positive, neutral, and negative labels as 1, 0, and -1. Multiplying the number of occurrences by the values and adding those together give a number that is divided by the total number of sentences. This is the sentiment score of the report which is compared to the stock movement. Since the neutral values are assigned to 0, there is no need to include that in the calculation. As an example, a report with 20 positive sentences, 120 neutral sentences, and 10 negative sentences would be calculated as the example below.

$$\frac{n_{positive} \cdot 1 + n_{negative} \cdot (-1)}{n_{positive} + n_{neutral} + n_{negative}} = \frac{20 \cdot 1 + 10 \cdot (-1)}{20 + 120 + 10} = 0.06\bar{6}\bar{6}$$

4.2.5 Visualizing and Correlation

The last step of the pipeline is the plotting of the sentiment for each report together with the return for the chosen period. This is visualized as a timeline from the oldest to the newest quarter, in other words as time-series data. To make the data comparable visually, the data has to be normalized. Visualization is not enough for systematic data comparison, therefore correlation has been added as a measure of the relation between the financial sentiment and the stock movement. The data from this process is stored as images and result files containing sentiment scores and stock returns. The correlation is a single measure for all of the reports and is therefore just collected as a single output number of the pipeline.

The sentiment scores are compared to the performance of the stock to detect the relation between the sentiment scores and the past and future returns. The next chapter describes the results of the process through illustrations and tables, which may help intuitive understanding of the methods.

5 Results

This chapter presents the results that are relevant to solve the initial problem statement and the research questions. The results are presented in a natural order based on the methodology of the work. One desired outcome of this chapter is that the same results can be achieved by the way the results are presented together with the presented methodology in the previous chapter. This chapter presents results from the text extraction, the sentiment prediction, and the relationship between stock performance and financial sentiment.

5.1 Text Extraction and Sentiment Prediction

The pipeline for extracting text and predicting financial sentiment generates output data for each file. In the output result files, the sentences are stored with the predicted sentiment for each sentence that has been extracted from the quarterly reports. Table 5.1 shows an example of how each result of a company is stored after extracting the sentences and the sentiment. Table 5.1 shows how the data are formatted and illustrates the typical sentiment prediction of the reports for the company Telenor ASA. Note that the order of reports is decided by the filenames but is sorted by date in the visualization pipeline.

filename	score	period
telenor_q1_2019.pdf	([(positive, 21), (neutral, 228), (negative, 1...	q1_2019
telenor_q1_2021.pdf	([(positive, 50), (negative, 90), (neutral, 22...	q1_2021
telenor_q1_2020.pdf	([(positive, 68), (neutral, 197), (negative, 4...	q1_2020
telenor_q4_2019.pdf	([(positive, 76), (negative, 52), (neutral, 28...	q4_2019
telenor_q4_2021.pdf	([(positive, 92), (negative, 86), (neutral, 40...	q4_2021
telenor_q4_2020.pdf	([(neutral, 239), (positive, 78), (negative, 6...	q4_2020

Table 5.1: Example of stored sentiment scores for reports. The data is structured by filename, sentiment score and period.

Table 5.1 does not show all the data generated in the score column, but Listing 7 below shows what is stored as the first element in the score column, which means the score of the first report of Telenor from the first quarter of 2019. The example does not show all the sentences, but a random selection.

Listing 7: Example of stored sentiment scores for a quarterly report

```
([('positive', 21), ('neutral', 228), ('negative', 12)],
 {'neutral': 4.8720574, 'positive': −4.1562567, 'negative': −3.451735},
  ('Total reported revenues were NOK 27.', 'neutral'),
  ('Currency adjusted gross profit declined by NOK 0.', 'negative'),
  ('NOK 0.', 'neutral'),
  ('EBITDA before other items was NOK 11.', 'neutral'),
  ('EBITDA margin of 40%, 1 percentage point below last year.', 'negative'),
  ('EBITDA declined by 3% on an organic basis.', 'negative'),
  ('On an organic basis, subscription and traffic revenues remained stable,
  while growth was 2.','positive'),
  ('Excluding Thailand, the organic subscription and traffic revenue growth
  was 1.', 'positive'))
```

The example in Listing 7 illustrates how sentences are predicted as negative, positive, and neutral. It also shows what sentences can look like after being extracted from PDF files. For most of the reports, neutral labels are the most common, but the number of negative and positive labels vary and is decisive for the overall sentiment of the report.

5.2 FinBERT Accuracy

For evaluation of the financial tone prediction, the accuracy can be calculated for each company. The accuracy shows the precision of the prediction of the BERT base model and the FinBERT model. The calculated accuracy is based on a manual evaluation of the predicted negative and positive labels and is calculated as presented in Chapter 2.5.2.

Table 5.2 shows that FinBERT delivers a higher accuracy than the base BERT model for financial tone prediction.

Company	FinBERT	BERT
Kahoot!	0.904	0.523
Schibsted	0.967	0.642
Telenor	0.975	0.667
Entra	0.833	0.658
Thon	0.943	0.561
Selvaag Bolig	0.967	0.600
Equinor	0.933	0.617
NEL	0.967	0.650
Aker BP	0.781	0.592
Average	0.919	0.612

Table 5.2: Accuracy of financial sentiment prediction for positive and negative sentences. FinBERT has a higher average accuracy than BERT.

In addition to the accuracy of correctly predicted positive and negative sentences, the sentences that are predicted as positive and negative were evaluated based on the completeness. Table 5.3 shows the fraction of complete sentences that were predicted to be positive or negative for BERT and FinBERT. Note that both models predict the same tokens, but FinBERT performs better in leaving broken sentences out of the prediction.

Company	FinBERT	BERT
Kahoot!	0.638	0.471
Schibsted	0.967	0.983
Telenor	0.750	0.800
Entra	0.967	0.966
Thon	0.874	0.912
Selvaag Bolig	0.850	0.874
Equinor	0.942	0.975
NEL	0.916	0.908
Aker BP	0.938	0.513
Average	0.871	0.822

Table 5.3: Percentage of complete sentences predicted as positive or negative by FinBERT and BERT

5.3 Performance Measurements

The following tables 5.4 and 5.5 show the performance measurements for FinBERT and BERT predictions of positive and negative tokens. The tables show that FinBERT performs better for all performance metrics, except precision.

Company	Accuracy	Misclassification	Precision	Recall	F1-score
Kahoot!	0.904	0.096	0.909	0.988	0.947
Schibsted	0.967	0.033	0.975	0.975	0.975
Telenor	0.975	0.025	0.970	0.984	0.977
Entra	0.833	0.167	0.916	0.854	0.884
Thon	0.943	0.057	0.917	0.978	0.946
Selvaag Bolig	0.967	0.033	0.967	0.967	0.967
Equinor	0.933	0.067	0.931	0.931	0.931
NEL	0.967	0.033	0.988	0.964	0.976
Aker BP	0.781	0.219	0.813	0.542	0.650
Average	0.919	0.081	0.932	0.910	0.917

Table 5.4: Performance of FinBERT prediction

Company	Accuracy	Misclassification	Precision	Recall	F1-score
Kahoot!	0.523	0.477	0.444	0.941	0.604
Schibsted	0.642	0.358	1.000	0.283	0.442
Telenor	0.667	0.333	1.000	0.111	0.200
Entra	0.658	0.342	1.000	0.281	0.438
Thon	0.561	0.439	1.000	0.167	0.286
Selvaag Bolig	0.600	0.400	1.000	0.238	0.385
Equinor	0.617	0.383	1.000	0.132	0.233
NEL	0.650	0.350	1.000	0.364	0.533
Aker BP	0.592	0.408	1.000	0.039	0.075
Average	0.612	0.388	0.938	0.284	0.355

Table 5.5: Performance of BERT prediction

5.4 Stock Performance and Predicted Sentiment

The illustrations in this section show the performance of the companies from when the report is published, plotted with the financial sentiment of each quarterly report. All numbers are normalized, and visualizations for all companies and return periods can be found in the Appendix.

5.4.1 One-Day Performance and Financial Sentiment

This subsection shows examples of visualization results for the computed one-day performance versus the predicted document financial sentiment score. A table with the correlation for all companies will be shown at the end.

Figure 5.1 shows the one-day performance of Telenor starting at the closing price on the last day before the quarterly report was published. The sentiment score is the calculated financial sentiment for the entire quarterly report. In other terms, the graph illustrates the relationship between predicted financial sentiment and the future one-day return.

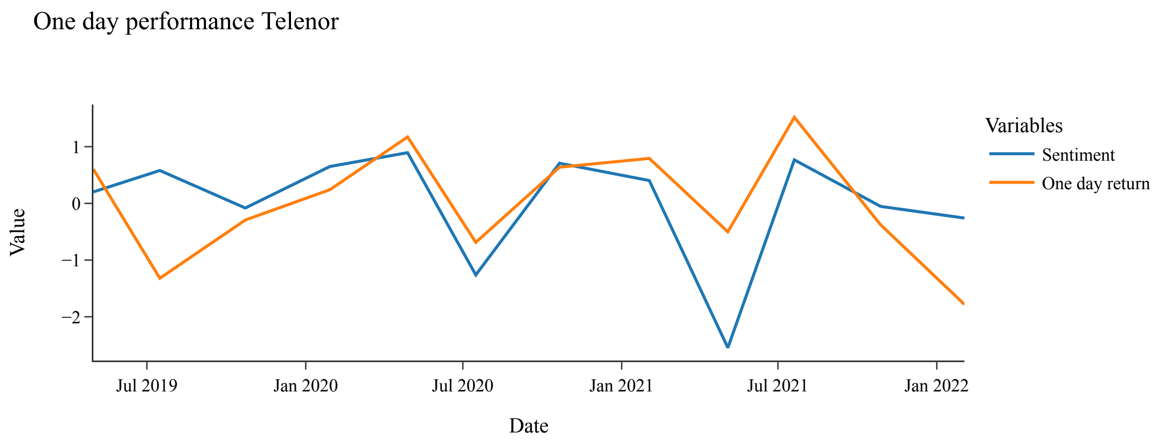


Figure 5.1: One-day performance and sentiment predicted for Telenor ASA

Figure 5.2 shows the one-day performance for Selvaag Bolig with the predicted report sentiment.

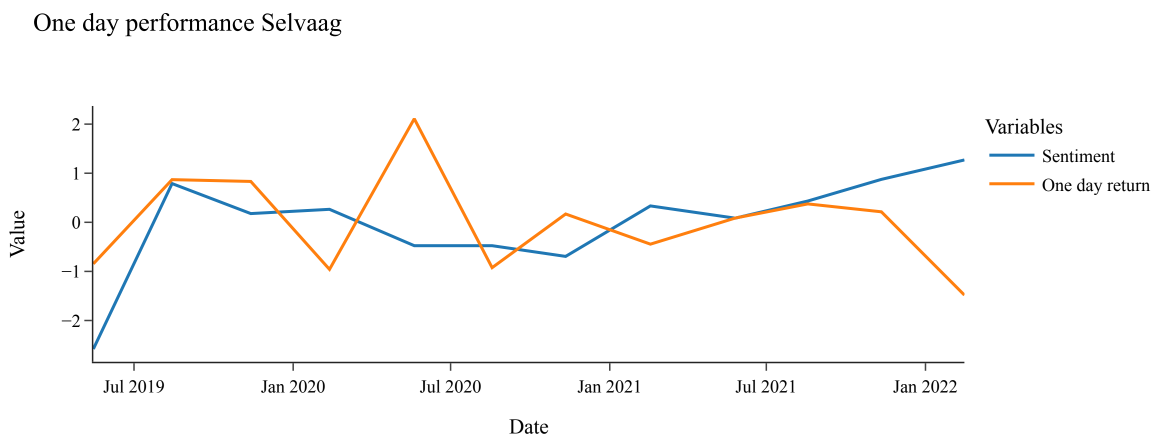


Figure 5.2: One-day performance and sentiment predicted for Selvaag Bolig ASA

The illustration in Figure 5.3 shows the one-day stock performance of Aker BP, which is a company in the energy sector.

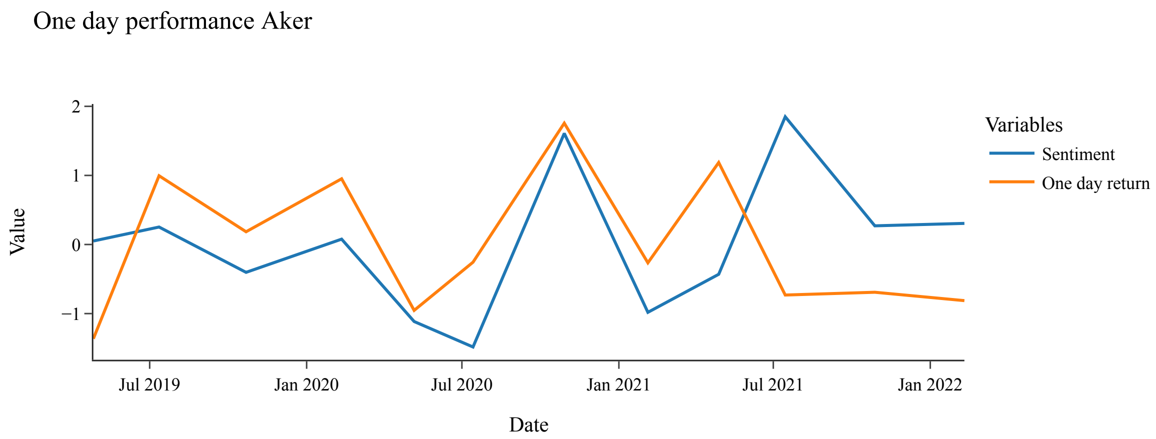


Figure 5.3: One-day performance and sentiment predicted for Aker BP ASA

Table 5.6 below shows the correlations between one-day returns and sentiment scores for all companies divided into industries and asset strategies. The correlation for the previously illustrated examples is in bold.

Technology		**Real Estate**		**Oil/Gas**	
Asset-light		*Asset-heavy*		*Asset-heavy*	
KAHOOT!	-0.195	ENTRA	-0.104	EQUINOR	-0.190
SCHIBSTED	-0.057	THON	-0.221	NEL	0.140
TELENOR	**0.468**	SELVAAG BOLIG	**0.037**	AKER BP	**0.219**
Sector Average	0.216	Sector Average	-0.096	Sector Average	0.056

Table 5.6: The correlation between one-day return and financial sentiment scores for all companies grouped by industry and asset strategy. The correlation is highest for Telenor and for technology companies on average.

The correlation is on average higher for the asset-light companies than for the asset-heavy companies when comparing one-day return and sentiment. The average correlation of all companies is 0.059 which is slightly higher than for one-week returns, but lower than the quarter return in the past. It can be observed that the relationship vary for the individual companies, which clearly affects the averaged figures.

5.4.2 One-Week Performance and Financial Sentiment

This subsection shows the visualization results for the computed one-week performance versus the predicted document financial sentiment score. A table with the correlation for all companies is provided at the end of the section.

The illustration in Figure 5.4 shows the one-week performance with the sentiment score for Schibsted.

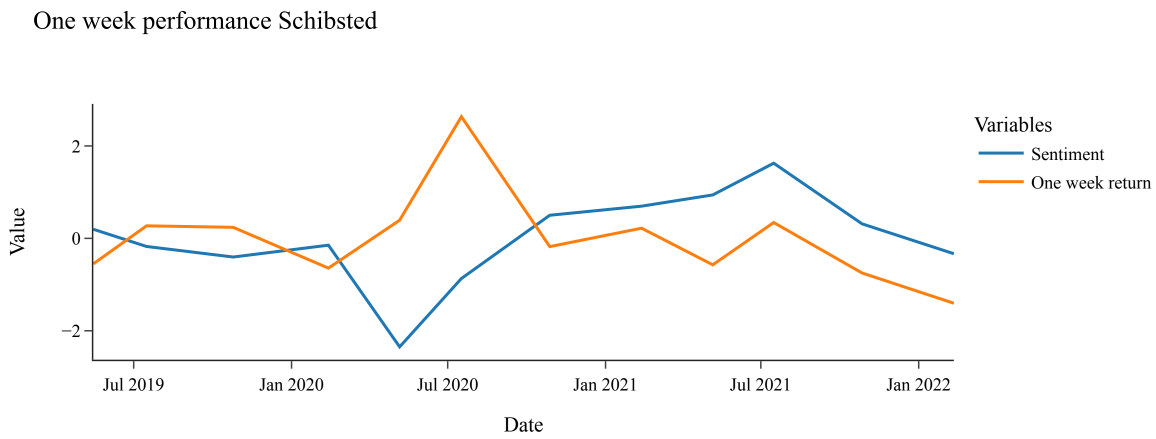


Figure 5.4: One-week performance and sentiment predicted for Schibsted ASA. The lines are following each other until spring 2020 which is when there was a lock down due to the Covid-19 pandemic.

Figure 5.5 shows the relationship between the predicted sentiment and the one-week return for Olav Thon Eiendomsselskap.

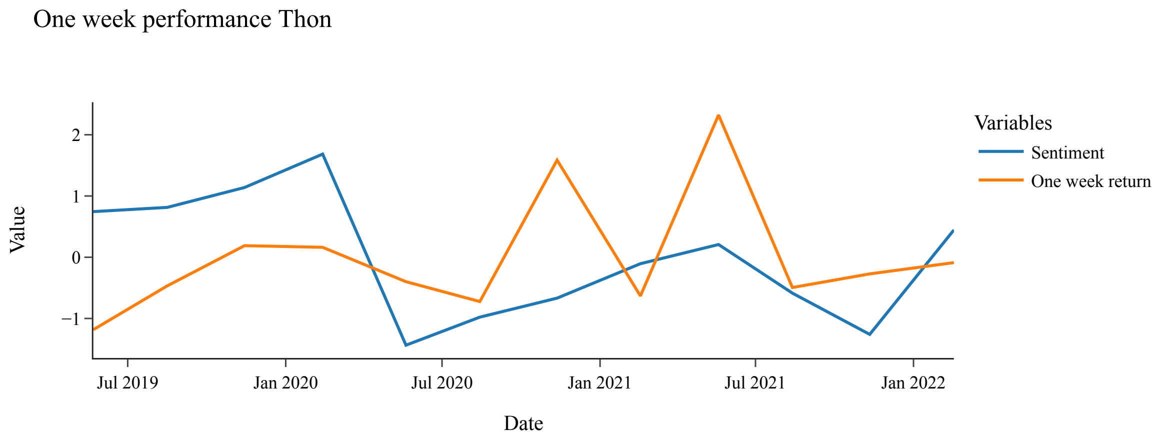


Figure 5.5: One-week performance and sentiment predicted for Olav Thon Eiendomsselskap ASA

The illustration in Figure 5.6 below shows a plot for the one-week performance and the predicted financial sentiment for NEL which is in the energy sector.

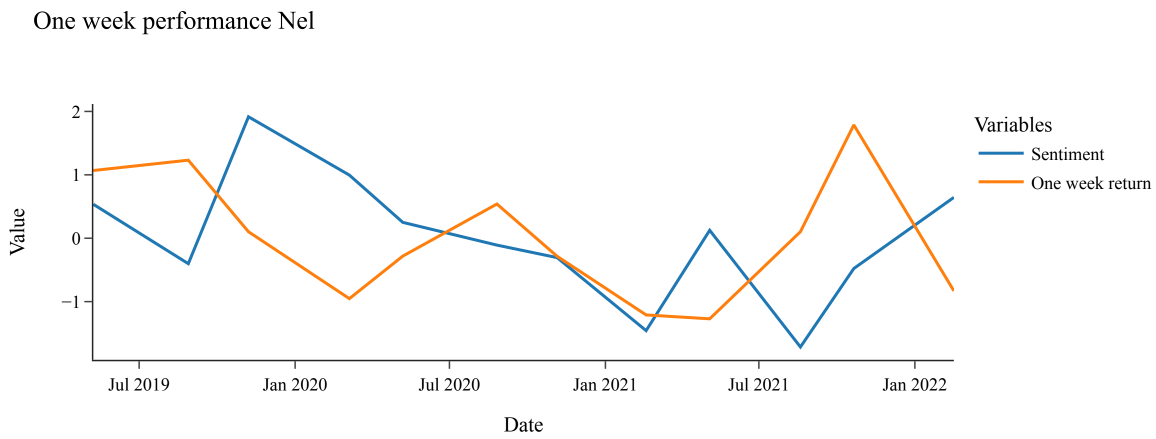


Figure 5.6: One-week performance and sentiment predicted for NEL ASA

Table 5.7 below shows the correlations between one-week returns and sentiment scores for all companies divided into industries and asset strategies. The correlation for the examples previously illustrated is in bold.

Technology		**Real Estate**		**Oil/Gas**	
Asset-light		*Asset-heavy*		*Asset-heavy*	
KAHOOT!	0.217	ENTRA	-0.367	EQUINOR	-0.049
SCHIBSTED	**-0.278**	THON	**0.054**	NEL	**-0.062**
TELENOR	0.250	SELVAAG BOLIG	0.042	AKER BP	-0.014
Sector Average	0.063	Sector Average	-0.090	Sector Average	-0.042

Table 5.7: The correlation between one-week return and sentiment scores for all companies

The correlation for each industry is low. One of the reasons is that the correlations vary from being positive to being negative, while others are close to zero. An overall correlation for all companies is -0.023. Telenor has the highest correlations with Kahoot! having the second highest. The correlation for Telenor can be categorized as a weak relationship based on the definition in Chapter 2.

5.4.3 One-Quarter Performance and Financial Sentiment

This section shows the visualization results for the computed performance for the reported quarter versus the predicted financial sentiment score of the report. A table with the correlation is presented for all companies at the end of the section.

The following three figures show the one-quarter past performance with the predicted sentiment of the quarterly report for Kahoot!, Entra, and Equinor.

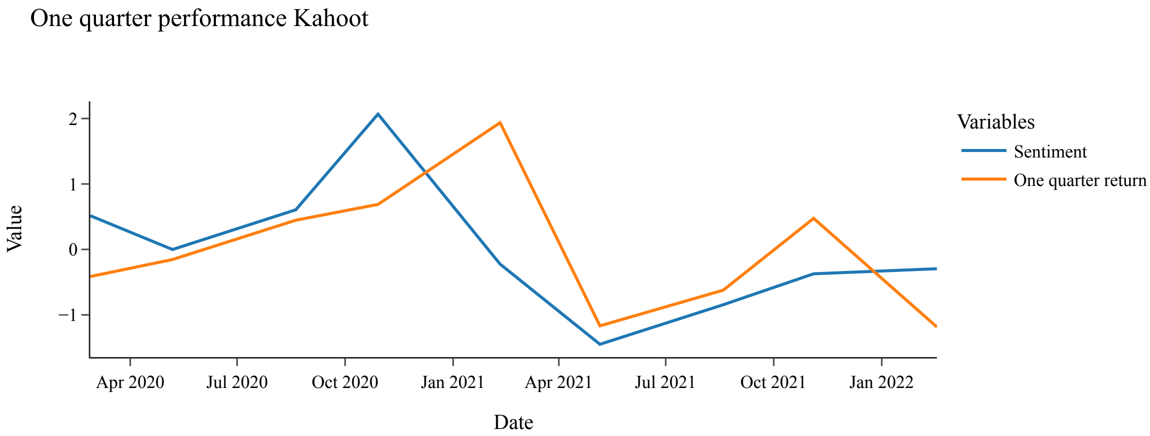


Figure 5.7: One-quarter performance and sentiment predicted for Kahoot! ASA

The next illustration in Figure 5.8 shows the relation between past quarter performance and predicted sentiment for Entra. Entra has the highest correlation for this relationship (0.500) which can be considered a moderate relationship.

Figure 5.8: One-quarter performance and sentiment predicted for Entra ASA

Equinor has a correlation of 0.365 and the relationship is illustrated in Figure 5.9 below. It is the highest correlation within the Oil/Gas sector of the included companies and the third highest among all the companies in the study.

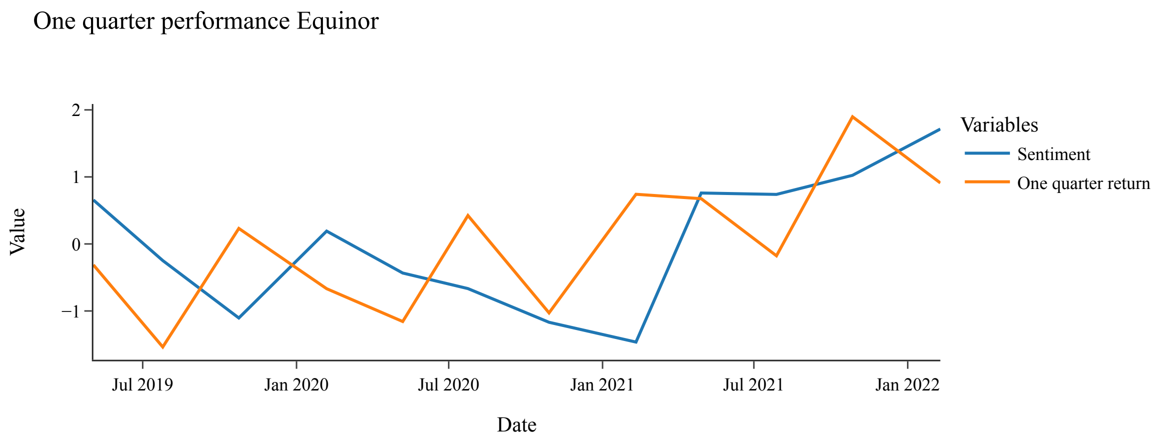


Figure 5.9: One-quarter performance and sentiment predicted for Equinor ASA

Table 5.8 shows the correlations between one-quarter returns for the reported period and the sentiment scores for all companies divided into industries and asset strategies. The correlations for the previously illustrated examples are in bold.

Technology		**Real Estate**		**Oil/Gas**	
Asset-light		*Asset-heavy*		*Asset-heavy*	
KAHOOT!	**0.430**	ENTRA	**0.500**	EQUINOR	**0.365**
SCHIBSTED	0.352	THON	0.175	NEL	0.263
TELENOR	0.043	SELVAAG BOLIG	0.022	AKER BP	-0.176
Sector Average	0.275	Sector Average	0.232	Sector Average	0.151

Table 5.8: The correlation between one-quarter past return and sentiment scores for all companies

For one-quarter past returns, the asset-light companies have the averagely highest correlation with the sentiment of the reports. On average, all companies have a correlation of 0.220 for the one-quarter return and the predicted sentiment from the report. This is higher than the one-week and one-day return against the sentiment. A weak to moderate relationship can be observed for Kahoot!, Schibsted, Entra, Equinor, and NEL. On average, a week relationship is observed for the asset-light technology companies.

5.4.4 One-to-One Correspondence

The relation presented in Table 5.9 shows the fraction of the corresponding negative and positive values for future returns and the sentiment of the report. A corresponding value would be counted if a binary relationship exists for positive and negative values.

Company	One-day	One-week	One-quarter
KAHOOT!	0.333	0.667	0.556
SCHIBSTED	0.417	0.417	0.583
TELENOR	0.667	0.500	0.500
ENTRA	0.417	0.417	0.750
THON	0.333	0.583	0.500
SELVAAG BOLIG	0.583	0.583	0.750
EQUINOR	0.500	0.333	0.667
NEL	0.417	0.500	0.583
AKER BP	0.667	0.750	0.500

Table 5.9: The fraction of the corresponding positive and negative values for return and sentiment

The results from one-to-one correspondence measures another type of relationship than correlation. For some companies this number is 75%. In other terms, e.g. the prediction of a negative or positive one-week stock movement would have been correctly predicted for 75% of the samples for Aker BP for the reported period.

5.4.5 Average Sentiment Scores

Another result to notice is the non-normalized sentiment scores and returns. This reveals individual differences in the language of the reports. Figure 5.10 and 5.11 show examples of one-day returns and the non-normalized sentiment scores for a selection of companies.

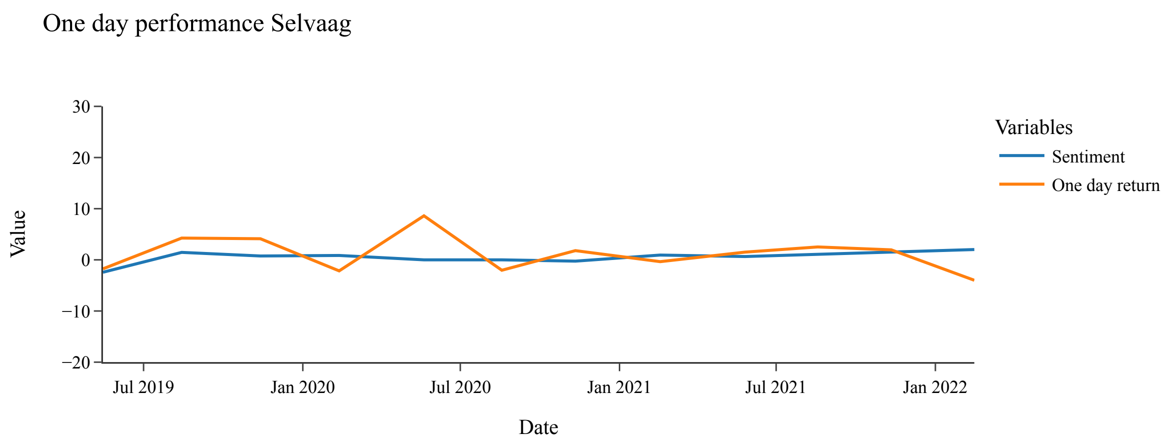


Figure 5.10: One-day performance and sentiment predicted for Selvaag Bolig ASA with non normalized figures

These give a good indication of the degree of neutral language of a report and that there are individual differences. The original sentiment score is multiplied by 100 to present the difference in sentiments of companies in a more intuitive way. As presented in Figure 5.10, Selvaag Bolig has a neutral predicted language in all reports, which can be seen more clearly when the figures are not normalized. On the other hand, Figure 5.11 shows that Kahoot! has a much more positive language in their reports.

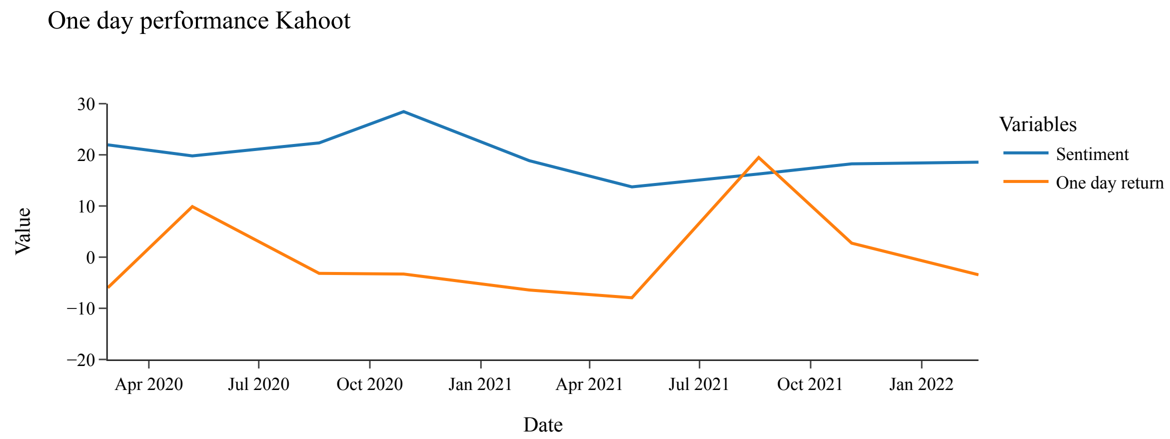


Figure 5.11: One-day performance and sentiment predicted for Kahoot! ASA with non normalized figures

Table 5.10 shows the average sentiment scores for all reports of the chosen companies. One can see from the result that the general BERT model predicts most of the tokens as negative, and therefore on average returns negative sentiment scores for all of the companies. The predicted score is not as positive as with the FinBERT model.

Company	**FinBERT**	**BERT**	**Return 2019-2021**
KAHOOT! ASA	19.797	22.058	228.6%*
SCHIBSTED ASA	14.303	-20.633	57.8%
TELENOR ASA	2.389	-23.222	-16.9%
ENTRA ASA	6.086	-13.368	73.4%
OLAV THON EIENDOMSSELSKAP ASA	1.374	-15.275	36.8%
SELVAAG BOLIG ASA	0.550	-7.169	28.5%
EQUINOR ASA	-3.217	-31.102	30.5%
NEL ASA	9.825	-14.575	221.5%
AKER BP ASA	-1.235	-19.516	26.5%

*Return from the start of the third quarter of 2019 until end of 2021.

Table 5.10: Average predicted sentiment scores and total returns for the period 2019-2021.

The correlation between FinBERT sentiment scores and total returns is 0.224, and the relationship is visualized in Figure 5.12. Furthermore, leaving Kahoot! out of the measurement gives a correlation of 0.550 which can be considered a moderate relationship. The correlation between the average sentiment predicted by the BERT base model and the total return is lower, at -0.078.

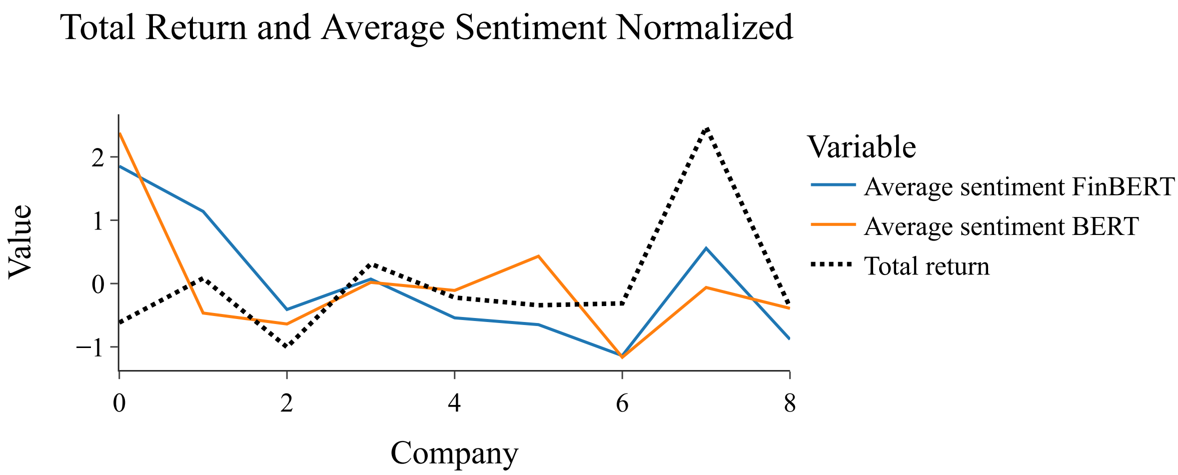


Figure 5.12: Total Return and average sentiment normalized

5.4.6 BERT Correlations

As an attempt to evaluate the performance of the pre-trained FinBERT model for financial sentiment prediction, a comparison of the correlations using the BERT base model is done. Table 5.11 below shows the correlation of the FinBERT model for all types of returns and the same for the BERT model.

Company	**FinBERT**	**BERT**	**FinBERT**	**BERT**	**FinBERT**	**BERT**
	one-day	one-day	one-week	one-week	past quarter	past quarter
Kahoot!	-0.212	-0.046	0.217	-0.078	0.430	0.246
Schibsted	-0.057	-0.409	-0.278	-0.558	0.352	-0.138
Telenor	0.468	0.135	0.250	-0.110	0.043	0.531
Entra	-0.104	0.391	-0.367	-0.030	0.500	0.057
Thon	-0.221	-0.176	0.054	0.062	0.175	-0.175
Selvaag Bolig	0.037	0.399	0.042	0.034	0.022	-0.720
Equinor	-0.190	-0.190	-0.049	-0.242	0.365	-0.061
NEL	0.140	0.050	-0.062	0.004	0.263	0.510
Aker BP	0.219	0.090	-0.014	-0.062	-0.176	-0.256
Average	0.009	0.027	-0.023	-0.109	0.219	-0.001

Table 5.11: Correlations for returns and sentiment

The table shows that the correlations for the BERT base model and the FinBERT model do not differ much on average for one-day and one-week stock return. The difference is more significant for the past quarter relationship. The predicted sentiment from the FinBERT model is on average more correlated (0.219) to the past quarter performance than the BERT base model (-0.001). Additionally, there are on average stronger positive correlations for the FinBERT model.

6 Discussion

The main goal of this study has been to analyze the potential of financial sentiment analysis of quarterly reports using the FinBERT model. Furthermore, the intention of the study is to demonstrate the use of domain-specific NLP for financial applications. Furthermore, the purpose has been to investigate the relationship between stock performance and sentiment scores, but also differences for specific asset strategies. Correlation has been the main measurement along with visualization plots as a foundation for the discussion.

This chapter presents a discussion based on the methods and results of this thesis, previous research, and the theoretical framework. The initial problem statement and research questions form the basis of the chapter alongside previous relevant research. Some of the results that deviate from previous research will also be discussed. The chapter will discuss in order of research questions and end with a general discussion. One of the purposes of this chapter is to discuss how the theory matches the findings of the methods and the results. Furthermore, discussing the methods and applied theory involves criticizing the choices made throughout the thesis work.

The main research question was; *Does the FinBERT model reveal the financial tone of a quarterly report and demonstrate potential for financial decision-making for Norwegian technology, real estate and energy companies?*

Furthermore, the main question was divided into three specific research questions, which were compiled based on suggestions for further work or identified gaps in the existing research. The three questions were constructed to discuss the following topics.

1. *The accuracy of the FinBERT model for financial sentiment prediction of quarterly reports*

2. *The relationship between predicted sentiment and stock movements*

3. *The dependency of asset strategy and industry for strength of relationship*

The questions will be discussed in the order of the methods relevant to answer the research questions. The first pipeline, which performs text extraction and sentiment analysis, is relevant to answer the first question. The second pipeline processes sentiment scores and compares them with past and future stock performance. This pipeline is generating results to answer the last two research questions.

6.1 Data

The data form the foundation of any study, and therefore, it is essential to address the issues with the data used in this study. The discussion of format, quantity, and quality of data is valuable for interpretation on the modeling performance, and discussion of research results.

In computer science, garbage-in-garbage-out (GIGO) is an expression for the quality of the output. In simple terms, it means that bad quality in the input data would produce a bad output quality. It is essential that the actual input data to the FinBERT model have the right format and represent the report well through the generated tokens. One has to be very critical to both the text extraction, which results in the input data for the sentiment prediction, and the predicted sentiment scores, which are the input data for the comparison. Furthermore, to discuss the quality of text extraction and predicted sentiment, one has to dive deeper into the input data.

6.1.1 Data Collection

Data are collected manually from the investor relations website for the selected companies. If possible, reports would have been downloaded directly through an application programming interface to increase efficiency and capacity. In this way, a large number of reports can be analyzed and additional industries can be included. The correlation between the average sentiment and the total return for all companies combined presented in Table 5.12, may imply that including more samples results in a weak overall relationship. Including more samples is a suggestion for future work for this study as it can add value to the research, such as the interpretation of correlations and the influence of industries and asset strategies. On the other hand, reliable in-depth analysis of quarterly reports demands a manageable number of reports, e.g. time-consuming tasks like sentiment labeling.

6.1.2 Data Format

The quarterly reports are the initial data and are essential in this study. All reports are PDF files, but the structure and layout vary. The large number of possible layout combinations makes automatic text extraction challenging. PDF is a difficult format to process as it is a flexible document format built specifically for readability and not for automated processing. The possibility of a wide range of layouts and elements combined with memory compression to decrease file size constitutes the challenge of extracting the desired text for the sentiment prediction. Filtering elements like tables, figures, and text is therefore a challenge task. Additionally, text elements can be presented as highlights, quotes, headlines, or comments, which makes the extraction even more difficult. The large number of financial figures is making the extraction of full sentences another challenge. However, the intention of this study is to perform financial sentiment analysis on quarterly reports, which are distributed as PDF files for public access. Additionally, extraction of text from financial PDF reports may contribute to research on financial analysis in general.

6.1.3 Typing Errors

Data entry tasks increase the risk of deficiency and faultiness in the data. The manual downloading and renaming of files can be a source of error. Furthermore, manual annotation of the report year and quarter increases the risk mentioned. The publication dates are also manually keyed as none of the PDF files contain dates. The use of filename-specific commands in the pipelines is an unwanted method, because any error in the filename would generate fallacious results. However, these methods were necessary for systematic data processing and visualization.

6.2 Text Extraction

Text extraction is an important process, as the generated output is the basis for sentiment prediction. The choice of extracting text with the most common font size may be a critical step. The extracted characters have the same font size, which means that the main text is extracted for documents with a consistent font style. This implies that highlighted sections with a different font size, which may contain valuable information, are not extracted. The text extraction of reports without style consistency will also suffer from this method. This is a clear weakness of the study and is recommended for future improvements. Additionally, the lack of cation shown in blank lines and white spaces influences the success of extraction. The method does, however, result in sentence extraction of all reports, and the application of sentiment prediction does not fail.

6.2.1 Compiling and Tokenization

The quality of the extracted text is essential for text compiling and tokenization. The detection of sentences by performing string matches, results in difficulties to recognize sentences that do not match the regular expression function. This is the case for sentences that do not end with a full point and, in some cases, results in longer tokens consisting of merged sentences. The model is, however, capable of predicting the correct sentiment of longer merged texts.

Table 5.3 presents the percentage of complete sentences that are predicted as positive or negative. The fraction of complete sentences is 87% for FinBERT and 82% for BERT. The main reason that the number is not higher is that Kahoot! has the lowest percentage of complete sentences, mainly due to the exclamation mark in the name, which is lowering the average. This resulted in extraction of sentences which ended at the exclamation mark because of the regular expressions. In addition, during the manual labeling, it was observed that the general BERT model predicted sentences ending with an exclamation mark as positive, which explains the low precision for Kahoot! presented in Table 5.5. However, the regular expression did not handle the full stop in financial figures, which resulted in a cutoff in sentences with decimal numbers and can be observed in Listing 7. Both problems could have been avoided by adding regular expressions for special cases. However, the accuracy of the prediction is still high for FinBERT, even if the extraction and cleaning of the text have limitations. FinBERT handles broken sentences well, as they are mostly predicted to be neutral. However, this increases the risk of splitting sentences that have positive or negative financial sentiment.

6.3 FinBERT Prediction

The first research question intends to investigate the performance of the FinBERT model for financial sentiment prediction of quarterly reports. Furthermore, it is desired to compare the results with the general BERT model to discuss the accuracy of the predictions with a financial BERT layer.

The illustration of the generated sentiment result in Listing 7 indicates a meaningful labeling of finance-related sentences. This representation of sentiment output indicates that predicting the financial sentiment was technically feasible. As stated in the previous subsection, the quality of the input data is critical for any modeling process. The illustration reveals that the issues are present in this study, as some of the tokens are not real sentences. On the other hand, these tokens are often predicted as neutral by the model and therefore do not change the balance of positive and negative sentiments.

6.3.1 Accuracy and Performance

As shown in Table 5.2, the accuracy is high for the positive and negative labels predicted with the FinBERT model. An initial investigation of the predicted sentiments reveals the challenge of text mining from PDF reports. On the other hand, the output of predicted labels seems promising, as tokens with meaningless text are often predicted as neutral. This may partly be the reason for the high number of neutral labels predicted for each report. Most figures in the result chapter are normalized, which means that the number of predicted positive and negative labels has an impact on the visual comparison. To quantify the true precision of sentiment prediction, the accuracy score is introduced.

To evaluate the accuracy of FinBERT prediction, a manual labeling is performed for all tokens extracted from the quarterly reports. The manual labeling is the only way to prove the accuracy of the FinBERT model and the performance compared to the BERT base model. This method reveals the true sentiment for each sentence of a quarterly report from a human perspective. Furthermore, true and false predictions can be detected and an accuracy score can be calculated, as presented in Table 5.2. The calculation of accuracy scores for FinBERT and BERT prediction reveals that FinBERT has a high accuracy and outperforms the BERT base model for financial sentiment prediction of quarterly reports. This emphasizes the results of previous research [23, 24], but is contradictory to the results of other studies [6], where the BERT base model had higher accuracy. A reason for this could be the latest upgrades in the FinBERT model in the last year. Moreover, the pre-training task of the FinBERT model on similar documents can be a reason for the high accuracy. In addition, the percentage of complete sentences is lower for BERT as presented in Table 5.3. Furthermore, BERT predictions result in high precision, but a low recall and F1-score. Elaborated, this means that for BERT the positive predictions are accurate, but the model fails for negative sentences, which may explain the negative sentiment scores presented in Table 5.10. FinBERT has high numbers for all performance metrics, indicating more balanced and robust prediction performance.

The FinBERT model has high accuracy in correctly predicted positive and negative sentences in a financial context, as shown in Table 5.2. This study also emphasizes the conclusion of Chen [23] that there is a disadvantage in using the base BERT model for a domain-specific case. This study shows that the FinBERT model performs better in predicting positive and negative sentences in the financial domain for quarterly reports. Furthermore, FinBERT predictions enable contextual understanding rather than a sum of sentiments for individual words.

6.4 Predicted Sentiment and Stock Performance

To discuss the second research question, the correlation between stock performance and report sentiment is calculated. In addition, the one-to-one correspondence has been computed for the binary relationship has been computed. For future stock performance, the one-day and one-week returns are computed. The past performance of the stock is the return of the stock for the reported quarter. This study compares the correlations for companies with different asset strategies from selected industries to answer the last research question.

6.4.1 Stock Market Complexity

Forecasting financial time series has been studied for many years, but the benefit of deep learning makes it possible to accomplish the task with less human effort [37]. It is the result of better models and cheaper computing, which yields more prediction capacity.

As a reminder, one should keep in mind the complexity of the stock market. The stock market is complex, and the variety of parameters is great. Expecting strong or moderate correlations between predicted sentiments and stock returns is optimistic. On the other hand, this study intends to dive into the complexity by searching for any observable relationships. The observed results are not convincing for stock buy recommendations. Vicari and Gaspari conclude in their study [37] that financial forecasting using deep learning models does not perform better than flipping a coin, which can be emphasized by the results of correlations between sentiments and one-day and one-week returns. However, monitoring companies through sentiment analysis of quarterly reports may result in more informed decision makers. An important task in investing is to reduce risk as much as possible. Using sentiment scores with weak or moderate relationships with the return of an investment decision may result in a failure, as there is a great possibility that the performance of the stock differs from the recommendation. However, this is normal behavior in the stock market, and investment decisions are mostly based on a larger number of factors and human intuition. This type of information could be a useful addition to the investment foundation, as it can provide information on the company's communication strategy and the historical correlation between future return and financial sentiment scores. Furthermore, expanding the number of reports, industries, and time intervals may equip financial analysts with valuable oversight.

6.4.2 Correlations

Defining a strong correlation between two variables can be difficult and can vary from one field to another. In this study, a strong correlation is defined as an absolute value of the correlation coefficient greater than 0.75. However, a strong correlation cannot be expected taking into account the complexity and the large number of factors that influence the performance of the stock. One still has to expect a number significant from zero to consider some sort of considerable linear relationship in addition to the interpretation of the graphs.

There is no clear sign of correlation of the sentiment and the one-week, nor one-day return after the report is published. There are, however, some interesting results when studying the companies individually. A correlation of 0.468 for Telenor ASA for the relation between the one-day return and financial sentiment is interesting. It is not large enough to be considered a moderate or strong correlation, but the number is clearly higher than any of the other companies, and studying the graph gives the impression of a relation between the numbers. This could be a coincidence, and there is no strong evidence to support the claim that they are strongly related, but it is an interesting result and may emphasize the complexity of the stock market and the characteristics of the different companies. However, this needs to be elaborated on in future work with more samples and time periods. The fact that performance and sentiment relation can be closer to some companies than others is not inconceivable to have some truth to it. Implementing individual stock monitoring based on financial sentiments may add value to decision-making for financial analysts.

Figure 5.12 visualizes the relationship between the total stock return and the predicted sentiment with BERT and FinBERT. Interestingly, the correlation between the FinBERT predictions and the total return is almost a weak relationship (0.224). This relationship increases to 0.550 when Kahoot! is left out of the comparison. Considering the lower accuracy of Kahoot! sentiment predictions, one can argue that it is reasonable to drop Kahoot! in the correlation. This chain of results implies that higher accuracy of financial sentiment prediction leads to a stronger correlation between sentiments and total returns.

6.4.2.1 COVID-19 Pandemic

Another interesting observation is the negative sentiment in the quarterly report of Schibsted after the shutdown due to the COVID-19 pandemic. Figure 5.4 shows the predicted negative sentiment for the report published in the spring 2020, but the return of the stock is strong after publication due to the general growth in the stock market. This emphasizes the complexity behind stock movements, and the impact of global incidents. The pandemic may have led to weaker relationships in this study due to lower financial stability, but it would have been easier to elaborate on with more reports over a longer period of time. This incident demonstrates the difficulty of evaluating predictions across business cycles and the necessity of more samples in this study.

6.4.3 FinBERT and BERT correlations

As presented in Table 5.11, the average correlation does not differ much between the BERT sentiment and the FinBERT sentiment for one-day and one-week performance.

However, there is a more significant difference for the past quarter performance. One could also tell from the individual correlations that the FinBERT model performs better in terms of higher correlation for specific companies, while the BERT model in general has negative average correlations for all time period comparisons. The BERT model does, however, yield weak correlations with the one-week returns for some companies such as Olav Thon Eiendomsselskap and Entra. A clear result is the consistency in the correlations for past quarter return and report sentiments using the FinBERT model, and the variation in the BERT model. This is another implications that the FinBERT model is performing better for stock performance prediction, but also for the prediction of financial sentiment of sentences. Taking into account the accuracy of predictions, this may strengthen the hypothesis of a weak financial relationship between financial sentiment of quarterly reports and stock performance.

6.4.4 Past Return

Another topic to address is whether companies communicate their actual performance for the quarter in terms of the financial sentiment of the report. Actual performance is simplified as the stock return for the reported quarter, and there are many other factors that can be considered to measure performance. Interestingly, the highest correlation, by far, is for the one-quarter return for the chosen quarter. This may imply that companies communicate in a way that reflects the performance of the company during the reported period to some extent. This is supported by the purpose of quarterly reporting, where performance can be a written evaluation of expectations. There are, however, individual differences as demonstrated through previous comparisons. This could imply that some companies are more honest about their performance than other companies, but it could also be a coincidence.

6.4.5 Average sentiment

Taking into account the results in Table 5.10, there is clearly a difference in the predicted language for the companies. Kahoot! ASA is an example of a company that has a more positive predicted financial language than other companies such as Selvaag Bolig ASA, as illustrated in Figure 5.11 and 5.10. This could imply that Selvaag Bolig ASA has a more neutral communication strategy, but, on the other hand, Kahoot! ASA has a higher correlation for the past quarter return (0.430) than Selvaag Bolig ASA (0.022), which means that Kahoot! is closer to communicating their actual performance. However, it is difficult to interpret these results, as the sentiment scores depend on the extraction method, which may have included more numbers for one report compared to another.

It can be seen from Figure 5.11 that for Kahoot! ASA, the model predicts a high positive sentiment score for all quarterly reports of the company, although the actual performance of the company is flat or negative for the analyzed quarters. However, the language of a quarterly report is expected to be professional and written in a neutral overall tone. Therefore, the results must be interpreted with care. This also implies that companies can publish reports in neutral language in periods of weak performance.

Overall, the Table 5.10 shows that companies with a higher average sentiment score also in general have higher total returns for the period. Another reason for the high sentiment score for Kahoot! may be that the company is young and inexperienced in financial reporting. In addition, the manual labeling resulted in a discovery of the chosen language style. Kahoot! has a more public-minded language, which contributed to the overall positive sentiment score.

6.4.6 One-to-One Correspondence

To alternatively assess the relationship, a binary measurement of the correct positive and negative prediction has been applied, as shown in Table 5.9. The table shows that the correspondence is the largest for the past return and the predicted sentiment. This is expected given that the correlation was highest for the same time interval. Aker BP has a correspondence of 75% for the one-week return, but on average the correspondence is around 50% for all companies. Therefore, no strong relationship can be established through correspondence, but rather a random directional change in the stock price compared to the sentiment value. Additionally, including more data points for measurement may result in a more reliable interpretation. However, the correspondence is higher for the performance of the quarter, which emphasizes the assumption that the predicted sentiment is more correlated with the performance in the past.

6.4.7 Asset Strategies and Industries

Table 5.6,5.7 and 5.8 show the correlations between past and future returns and predicted sentiment. The companies are divided into asset strategies and industries, and the averaged correlations are influenced by an oscillation in positive and negative values for each industry. The study by Zhou et al. [22] that suggested a larger prediction gain for asset-heavy industries cannot be supported by this study. In fact, the correlation between stock performance and predicted sentiment is higher for asset-light companies. This is valid regardless of the type of return used in the calculation. However, the previous study [22] did not include sentiment scores for the prediction, which may be the reason for the deviation. This may imply that the predictive gain is higher for asset-heavy companies when quarterly figures are used, but not for financial sentiments. Additionally, Table 5.9 shows that the correspondence is higher for asset-heavy companies and, therefore, the definition of predicting gain is decisive for the conclusion. Furthermore, the groundwork for the interpretation is weak due to the limited number of companies included in the thesis. However, the conclusion based on the results of this research is that the correlations vary independently of asset strategy and industry. Additionally, the strength of relationship is not higher than weak to moderate for any industry. On the other hand, the slightly higher correlation for asset-light companies for all time periods may indicate that companies with fewer assets emphasize communicating in a way that is recognizable to the business. On the other hand, asset-heavy companies may emphasize quarterly figures, which may be one of the reasons for the result of the study by Zhou et al. [22].

The result can be seen in light of another study [48] that finds that the structure of asset-light companies matters in reducing information uncertainty. Furthermore, the study [48] points out that investors overestimate the underlying risk of asset-heavy companies, which can be supported by this thesis considering Equinor and Aker BP being the only companies with a negative average financial sentiment, despite positive total returns.

6.4.8 Pipeline

As text extraction is a challenging task, this study demonstrates the technical feasibility with the use of Python tools and constructed ETL pipelines. Despite technical challenges, the pipeline structure contributes higher credibility to the study, because the pipelines produce data in a systematic manner and simplify the testing. This results in comparable output data and reduces the risk of computer variable conflicts. To ensure the credibility of the pipeline, each process was tested in an interactive programming notebook. In addition, all output files were stored with a version control so that changes in pipeline processes would not overwrite existing files. An advantage of using the pipeline structure is the possibility to do new experiences by changing parameters and run the code by one command, e.g. performing the same comparison for different time intervals. This thesis demonstrates the potential value of pipelines for financial analysis and monitoring applications, as they can be easily applied to different types of financial reports and stock data.

6.5 Pitfalls

In this study, multiple choices are made that probably have a considerable effect on the results and the credibility of the study. Therefore, the potential pitfalls are presented to caution the reader of critical methods and interpretations.

6.5.1 Normalization

Normalizing the data changes the basis for visual comparisons and interpretation of the data. As the scale changes, it is easy to consider the relationship based on the lines in the graph. However, it is difficult to draw further conclusions, as the scale is changing and the actual sentiment scores are not presented. However, those scores have been discussed based on the average sentiment scores on non-normalized data.

It is important to emphasize that the sentiment is exaggerated, as it is mainly neutral language in reports when applying the FinBERT model for prediction. Normalizing the results gives variation, which can be a coincidence. Without normalization, the sentiment graph would be close to zero for some of the companies, as they sometimes have fewer than ten sentences in total predicted as positive or negative.

6.5.2 Visualization Optimism

It is optimistic to consider visualization of only one measurement along with the stock price and to expect to see any relationship. The complexity of the stock market typically requires the inclusion of other measurements to provide expectations of higher correlation. In addition, more informative graphs are required to enable the discussion of relationships. However, the consideration of only two simple measurements is an efficient way of discussing the relationship of the two variables and a basis for a discussion of the initial problem statement of the thesis.

6.5.3 Correlation with Care

Although using correlation in this thesis, one should be careful with using correlation. It is one of the most widely used methods to measure relationships, but it can easily be misinterpreted. A weak linear relationship does not disqualify other kinds of relationships. On the other hand, a strong linear correlation does not necessarily mean that there is causation. In this study, linear correlation was used because it is a common measurement in studies comparing stock performance and sentiment scores. It is the main measurement of the studies to which this thesis refers [23, 19, 22].

Another reason to show care when interpreting correlations is the vulnerability to extreme values, which was demonstrated by the change in correlation when Kahoot! was left out. The correlation coefficient between average sentiment and total return changed from 0.224 to 0.550. This may be an important point for this work, as stock performance is influenced by a wide range of factors, such that there will be a huge reason for a stock decrease or increase that is not presented in the quarterly reports.

6.5.4 Calculation of Return

The stock price data are downloaded with the Python tool yfinance, which is built on the reliable Yahoo! finance API. Therefore, one can assume sufficient quality in the data. The downloaded open and close prices are the basis of the return computations. However, the type of return computation is simple and does not implement dividend, which may have an effect on the stock performance. Furthermore, opening and closing prices can, in specific cases, contain exaggerated values and result in misleading results. However, the method is consistent and can be expanded in future improvements if more complexity is desired. Additionally, the method is technically feasible and is effective in quantifying the performance for different time intervals.

6.5.5 Overall Report Sentiment

An important step of the study is to predict the financial sentiment scores for the reports. This is the basis for the comparison, and the method can be criticized. Transforming the sentiment labels to -1 for negative, 0 for neutral and 1 for positive is a simple way to explain the sentiment of a financial report. This choice of method results in a score that does not take into account the predicted values before applying the softmax function for label assignment. However, one can argue that the method represents the document to an adequate degree, as one of the purposes of the study is to predict the sentiments of the individual sentences. In addition, the high accuracy that was observed for FinBERT predictions, and the purpose of sentiment analysis to assign labels, may justify the choice.

6.6 Research Contributions

This thesis demonstrates a one-language workflow with Python for text extraction and cleaning of PDF files. The study also demonstrates the technical viability, but also the challenges, of sentiment analysis of PDF documents. The benefit of developing ETL pipelines for data-related tasks is also demonstrated throughout the study.

The challenge of extracting and cleaning text from PDF documents has been documented in previous research [49]. This thesis encounters the challenge of variation in the layout and elements in quarterly reports. The use of PDFMiner and regular expression results in 87% complete sentences predicted as positive or negative with FinBERT.

Furthermore, financial sentiments are predicted from the cleaned text, which results in an accuracy of 91.9% with FinBERT, which is higher than for the generalized BERT model (61.2%). The accuracy of FinBERT for quarterly reports is higher than the official documentation of the model (88.7%) [15] and in previous research (89%) [6]. However, these results are from before the latest upgrade of the model in 2021. The BERT model performs worse than in previous research in which accuracies between 74% [23] and 93% [6] have been reported. One of the reasons for low accuracy for the generalized BERT model in this thesis is that BERT has problems handling incomplete sentences which obviously increase the number of wrongly predicted sentences.

This study does not discover a strong relationship between predicted financial sentiment and future stock performance. This result differs from existing research in which a strong relationship was reported between the text of annual reports and the abnormal return [21]. No measurable effect on the stock market is identified to the same extent as studies that include sentiments from the news headlines [19]. One of the reasons for this may be the general language of quarterly reports. Additionally, the complexity of a movement in a stock can be another reason for the weak relationship. This is also supported by previous research [19], suggesting that more features should be included in future research. However, a weak or moderate relationship between individual companies may be relevant to a financial analyst. Furthermore, binary correspondence may be valuable information for investment decision-making. In addition, it was discovered that higher prediction accuracy leads to a stronger correlation between stock return and financial sentiment.

Accurate prediction of sentences does not imply accurate and representative text extraction. There is also the possibility that individual companies have a tendency to write reports with a more positive sentiment to not scare investors. This study reveals a variation in the average financial sentiment of the selected companies. This result can give insight into the communication strategy of a company.

Despite a weak correlation, individual relationships can add value to the monitoring of specific stocks. Previous research [22] also finds that forecast performance with quarterly data is not universal and varies for individual companies and industries. Additionally, this study finds that the predicted sentiment of quarterly reports is more correlated with past performance than future one-day and one-week performance. This is reasonable considering that the intention of the quarterly report is to report on the past quarter. In addition, the relationship is stronger for longer time intervals, which is also demonstrated in the literature [24].

Furthermore, the sentiment predicted for asset-light technology companies tends to be more correlated with stock performance than for asset-heavy companies, which differs from the study by Zhou et al. [22]. However, another study [48] reveals that asset-heavy companies have underlying risk that causes overestimated negative sentiment. This may be supported by this thesis, as less information uncertainty can be the reason for stronger correlation between stock performance and financial sentiment for asset-light companies. However, because the difference is small and the number of industries and companies is limited in this thesis, the result must be interpreted with caution.

7 Conclusion

This chapter gives a conclusion based on the discussion of the main results of the thesis. In addition, it gives an answer to what this research has contributed to the initial problem statement. The chapter ends with suggestions for future work on the study.

Despite all the weaknesses of this thesis, the development of NLP techniques emphasizes the potential for applications in the financial industry. This study demonstrates the use of pre-trained language models for financial reports, and it is assumed that the gap in research gives clear motivation for further work in the particular research area.

7.1 Main Results

As there are a large number of specific results that can be presented in a study that generates data, only the few major results relevant to the initial research questions are presented in the list below.

- FinBERT predicts the financial sentiment of quarterly reports with a higher accuracy (91.9%) than the generalized BERT model (61.2%).
- There is no strong correlation between future stock performance and the predicted sentiment of the quarterly reports on average, but a weak to moderate relationship for individual stocks.
- The average correlation between stock performance and predicted financial sentiment is stronger for the return of the reported quarter (0.220).
- A weak correlation is found between the average predicted sentiment of the quarterly reports and the total return for the same period.
- This study demonstrates that a higher accuracy in the prediction of financial sentiment leads to a stronger correlation between the sentiment of a quarterly report and the total return of a stock.
- The correlation between the stock return and the predicted sentiment is stronger for asset-light technology companies, but the sentiment of asset-heavy companies has a higher one-to-one binary correspondence for positive and negative values of

stock movements and predicted sentiment.

7.2 Future Work

This study demonstrates a workflow for financial sentiment analysis of quarterly reports and the relation to past and future stock performance. However, financial forecasting is difficult, as concluded in previous research [37]. Considering the complexity of stock movements, future research is recommended to include additional text sources to measure the relationship between financial sentiments and stock movements. However, the importance of the GIGO concept should be considered in future work, as this study can be improved by better text extraction and text cleaning. The following list provides suggestions for future work on the study.

- Improving the quality of text extraction and cleaning should be the focus of future work in this study. Furthermore, grouping and predicting text for different topics, such as sustainability, would be an interesting extension for further domain insight.
- The number of reports, companies, and industries should be increased to provide more reliable results for the relationship between stock performance and sentiment scores. An API endpoint for automatic download of reports and dates would be beneficial for this purpose and for a decision-support prototype.
- The inclusion of additional text sources can be a potential extension of the study under the condition of improved text extraction and text cleaning.

References

[1] A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, "Big data monetization throughout big data value chain: a comprehensive review," *Journal of Big Data*, vol. 7, no. 1, p. 3, 2020. [Online]. Available: https://doi.org/10.1186/s40537-019-0281-5

[2] M. M. Hasan, J. Popp, and J. Oláh, "Current landscape and influence of big data on finance," *Journal of Big Data*, vol. 7, no. 1, p. 21, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00291-z

[3] X. Cheng, S. Liu, X. Sun, Z. Wang, H. Zhou, Y. Shao, and H. Shen, "Combating emerging financial risks in the big data era: A perspective review," *Fundamental Research*, vol. 1, no. 5, pp. 595–606, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667325821001722

[4] S. Khan and M. R. Rabbani, "Artificial intelligence and nlp -based chatbot for islamic banking and finance," *INTERNATIONAL JOURNAL OF INFORMATION RETRIEVAL RESEARCH*, vol. 11, no. 3, pp. 65–77, 2021.

[5] C. Masson and P. Paroubek, "Nlp analytics in finance with dore: A french 250m tokens corpus of corporate annual reports," in *LREC*, 2020.

[6] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131 662–131 682, 2020.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," p. arXiv:1810.04805, October 01, 2018 2018. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D

[8] H. M, G. E.A., V. K. Menon, and S. K.P., "Nse stock market prediction using deep-learning models," *Procedia Computer Science*, vol. 132, pp. 1351–1362, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918307828

[9] M. Mailagaha Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Systems with Applications*, vol. 197, p. 116659, 02 2022.

[10] K. Mishev, A. Gjorgjevikj, R. Stojanov, I. Mishkovski, I. Vodenska, L. Chitkushev, and D. Trajanov, "Performance evaluation of word and sentence embeddings for finance headlines sentiment analysis," in *ICT Innovations 2019. Big Data Processing and Mining*, S. Gievska and G. Madjarov, Eds. Cham: Springer International Publishing, 2019, pp. 161–172.

[11] R. Gupta and M. Chen, "Sentiment analysis for stock price prediction," in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2020, pp. 213–218.

[12] K. C. Butler and S. Malaikah, "Efficiency and inefficiency in thinly traded stock markets: Kuwait and saudi arabia," *Journal of Banking & Finance*, vol. 16, no. 1, pp. 197–210, 1992, special Issue on European Capital Markets. [Online]. Available: https://www.sciencedirect.com/science/article/pii/037842669290085E

[13] L. A. Gallagher and M. P. Taylor, "Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks," *Southern Economic Journal*, vol. 69, no. 2, pp. 345–362, 2002.

[14] B. Qian and K. Rasheed, "Stock market prediction with multiple classifiers," *Applied Intelligence*, vol. 26, no. 1, pp. 25–33, 2007.

[15] Y. Yang, M. C. S. Uy, and A. Huang, "Finbert: A pretrained language model for financial communications," 2020.

[16] N. Rouf, M. B. Malik, T. Arif, S. Sharma, S. Singh, S. Aich, and H.-C. Kim, "Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions," *Electronics*, vol. 10, no. 21, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/21/2717

[17] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, p. 102–107, mar 2016. [Online]. Available: https://doi.org/10.1109/MIS.2016.31

[18] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187775031100007X

[19] L. Nemes and A. Kiss, "Prediction of stock values changes using sentiment analysis of stock news headlines," *Journal of information and telecommunication (Print)*, vol. 5, no. 3, pp. 375–394, 2021.

[20] P. Sonkiya, V. Bajpai, and A. Bansal, "Stock price prediction using bert and gan," 2021. [Online]. Available: https://arxiv.org/abs/2107.09055

[21] P. Hájek, "Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns," *Neural Computing and Applications*, vol. 29, no. 7, pp. 343–358, 2018. [Online]. Available: https://doi.org/10.1007/s00521-017-3194-2

[22] H. Zhou, W. O. Maneesoonthorn, and X. B. Chen, "The predictive ability of quarterly financial statements," *International Journal of Financial Studies*, vol. 9, no. 3, p. 50, 2021. [Online]. Available: https://www.mdpi.com/2227-7072/9/3/50

[23] Q. Chen, "Stock movement prediction with financial news using contextualized embedding from bert," 2021. [Online]. Available: https://arxiv.org/abs/2107.08721

[24] C. M. Liapis, A. Karanikola, and S. Kotsiantis, "A multi-method survey on the use of sentiment analysis in multivariate financial time series forecasting," *Entropy*, vol. 23, no. 12, p. 1603, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/12/1603

[25] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017, multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885617301191

[26] W. C. Freund, M. Larrain, and M. S. Pagano, "Market efficiency before and after the introduction of electronic trading at the toronto stock exchange," *Review of Financial Economics*, vol. 6, no. 1, pp. 29–56, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1058330097900136

[27] R. Næs, J. A. Skjeltorp, and B. A. Ødegaard, *What factors affect the Oslo Stock Exchange?* Norges Bank, 2009.

[28] "https://www.oslobors.no/ob_eng/oslo-boers/regulations/oslo-rule-book-ii-issuer-rules," 2021. [Online]. Available: https://www.oslobors.no/ob_eng/Oslo-Boers/Regulations/Oslo-Rule-Book-II-Issuer-Rules

[29] "Ias 34 interim financial reporting," 2021. [Online]. Available: https://www.ifrs.org/issued-standards/list-of-standards/ias-34-interim-financial-reporting/#IAS34_1

[30] A. Raj, J. Bosch, H. H. Olsson, and T. J. Wang, "Modelling data pipelines," *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (Seaa 2020)*, pp. 13–20, 2020, bs2jj Times Cited:0 Cited References Count:20 EUROMICRO Conference Proceedings. [Online]. Available: <GotoISI>://WOS:000702094100003

[31] J. Fan, "Text segmentation of consumer magazines in pdf format," in *11TH INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION (ICDAR 2011)*, ser. Proceedings of the International Conference on Document Analysis and Recognition. Chinese Acad Sci; FOUNDER; NSFC; FUJITSU; Hanvon Technol; Green Apple Data Ctr; ABBYY; Raytheon BBN Technologies; Nuance; A2iA; ITESOFT; Int Assoc Pattern Recognit, TC10 Graph Recognit & TC11 Reading Syst; CAA; Tsinghua Univ; CASIA; HITACHI, 2011, pp. 794–798, 11th International Conference on Document Analysis and Recognition (ICDAR), Beijing, PEOPLES R CHINA, SEP 18-21, 2011.

[32] H. Bast and C. Korzen, "A benchmark and evaluation for text extraction from pdf," in *2017 ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES (JCDL 2017)*, ser. ACM-IEEE Joint Conference on Digital Libraries JCDL. IEEE; ACM; Univ Toronto, Fac Informat; Univ Toronto, Libraries; Libraries Bloomington; Elsevier; DLF; VirginiaTech, 2017, pp. 99–108, aCM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, CANADA, JUN 19-23, 2017.

[33] "pdfminer.six," 2020. [Online]. Available: https://github.com/pdfminer/pdfminer.six

[34] M. S. U. Miah, J. Sulaiman, T. B. Sarwar, A. Naseer, F. Ashraf, K. Z. Zamli, and R. Jose, "Sentence boundary extraction from scientific literature of electric double layer capacitor domain: Tools and techniques," *APPLIED SCIENCES-BASEL*, vol. 12, no. 3, FEB 2022.

[35] K. S. Nugroho, A. Y. Sukmadewa, and N. Yudistira, “Large-scale news classification using bert language model: Spark nlp approach,” in *6th International Conference on Sustainable Information Engineering and Technology 2021*, ser. SIET ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 240–246. [Online]. Available: https://doi.org/10.1145/3479645.3479658

[36] D. Martín-Gutiérrez, G. Hernández-Peñaloza, A. B. Hernández, A. Lozano-Diez, and F. Álvarez, “A deep learning approach for robust detection of bots in twitter using transformers,” *IEEE Access*, vol. 9, pp. 54 591–54 601, 2021.

[37] M. Vicari and M. Gaspari, “Analysis of news sentiments using natural language processing and deep learning,” *AI & society*, vol. 36, no. 3, pp. 931–937, 2020.

[38] Y. Luo and L. Zhou, “Textual tone in corporate financial disclosures: a survey of the literature,” *International Journal of Disclosure and Governance*, vol. 17, no. 2, pp. 101–110, 2020. [Online]. Available: https://doi.org/10.1057/s41310-020-00077-y

[39] A. M. Hashmi, F. Qayyum, and M. T. Afzal, “Insights to the state-of-the-art pdf extraction techniques,” *Ipsi Bgd Transactions on Internet Research*, vol. 16, no. 1, pp. 60–67, 2020, hashmi, Ahmer Maqsood Qayyum, Faiza Afzal, Muhammad Tanvir Si. [Online]. Available: <GotoISI>://WOS:000573367200009

[40] Y. Yang, M. C. S. Uy, and A. Huang, “Finbert: A pretrained language model for financial communications,” 2021. [Online]. Available: https://github.com/yya518/FinBERT

[41] J. A. Thompson, J. Tan, and C. S. Greene, “Cross-platform normalization of microarray and rna-seq data for machine learning applications,” *PeerJ*, vol. 4, p. e1621, Jan. 2016. [Online]. Available: https://doi.org/10.7717/peerj.1621

[42] W. McKinney *et al.*, “pandas: a foundational python library for data analysis and statistics,” *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.

[43] P. McClure, “Correlation statistics: Review of the basics and some common pitfalls,” *Journal of Hand Therapy*, vol. 18, no. 3, pp. 378–380, 2005.

[44] S. B. Caudill, M. Ayuso, and M. Guillén, “Fraud detection using a multinomial logit model with missing information,” *Journal of Risk and Insurance*, vol. 72, no. 4, pp. 539–550, 2005.

[45] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[46] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, “Thresholding classifiers to maximize f1 score,” *arXiv preprint arXiv:1402.1892*, 2014.

[47] J. Singh, B. McCann, R. Socher, and C. Xiong, “Bert is not an interlingua and the bias of tokenization,” in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, 2019, pp. 47–55.

[48] C. Poretti and C. Y. Heo, “Asset-light strategies and stock market reactions to covid-19’s pandemic announcement: The case of hospitality firms,” *Tourism economics*, p. 13548166211005198, 2021.

[49] H. Bast and C. Korzen, “A benchmark and evaluation for text extraction from pdf,” in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2017, pp. 1–10.

Appendix

A0.1 One-day Return

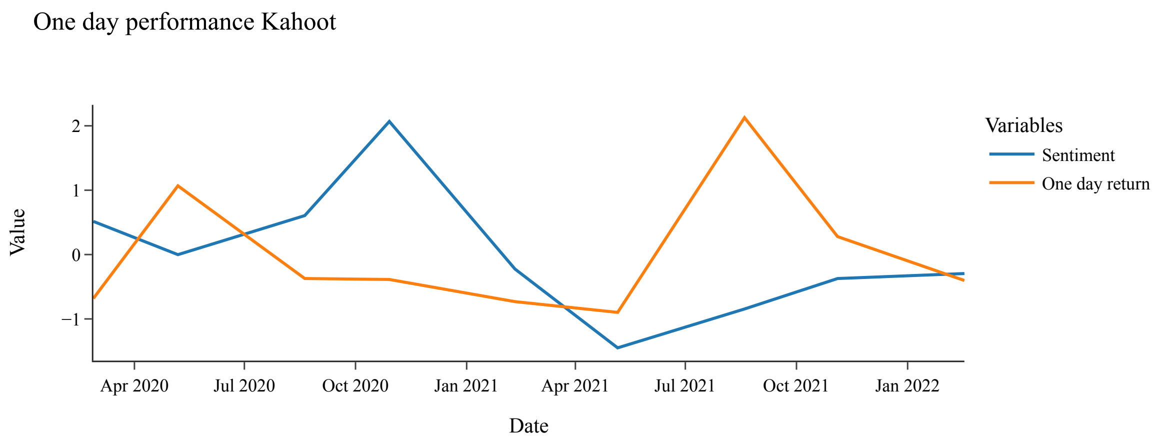


Figure A0.1: One-day performance and sentiment predicted for Kahoot ASA

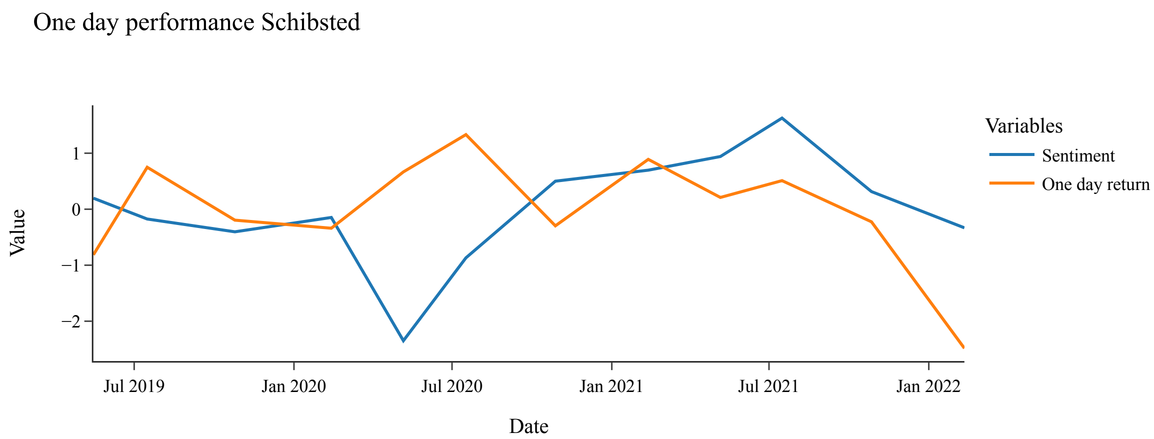


Figure A0.2: One-day performance and sentiment predicted for Schibsted ASA

A0.2 One-week Return

A0.3 One-quarter Return

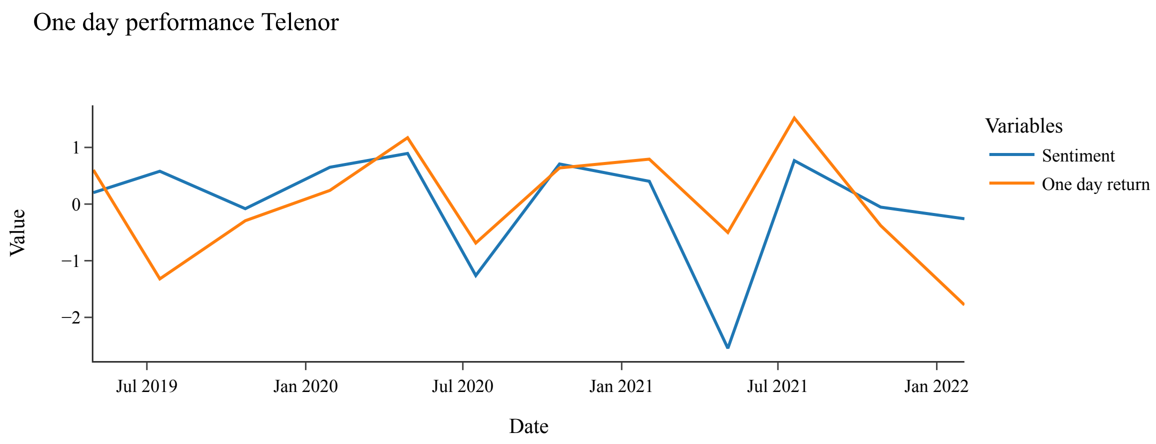


Figure A0.3: One-day performance and sentiment predicted for Telenor ASA

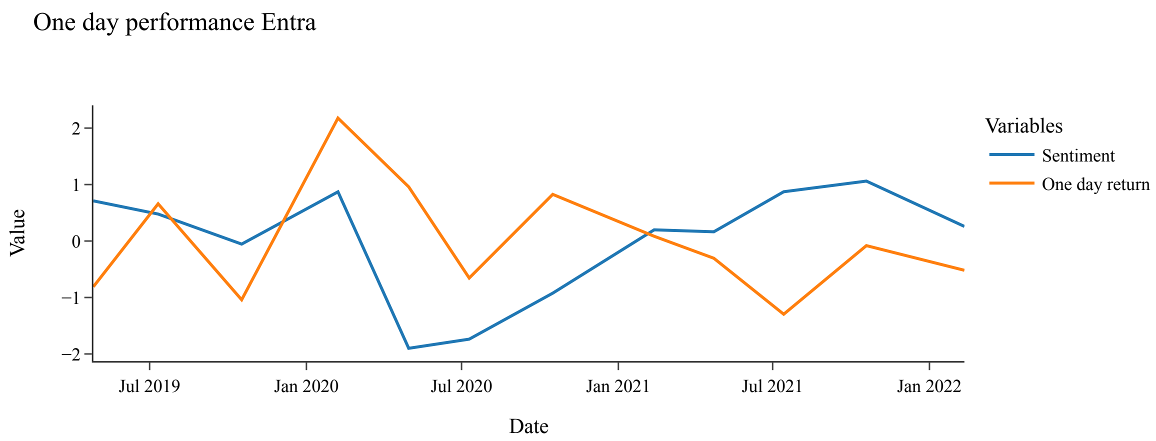


Figure A0.4: One-day performance and sentiment predicted for Entra ASA

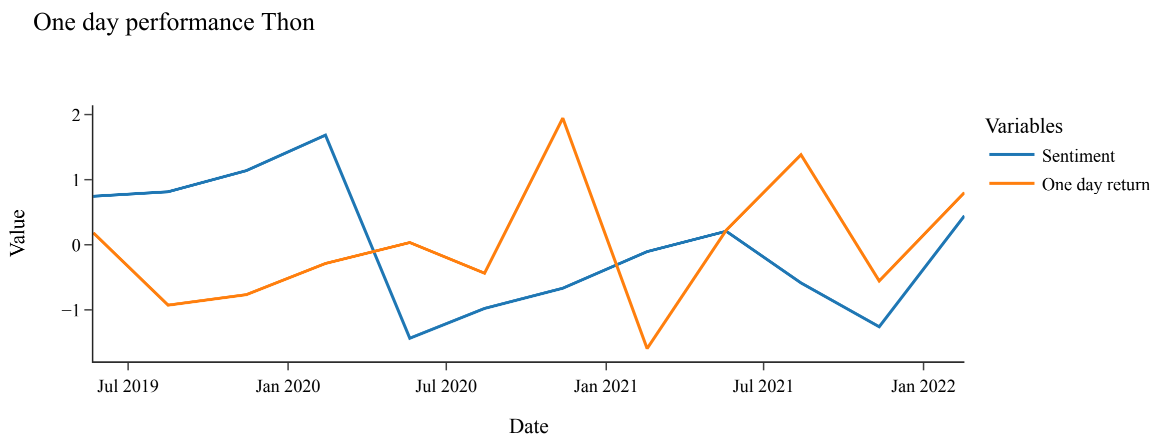


Figure A0.5: One-day performance and sentiment predicted for Olav Thon Eiendomsselskap ASA

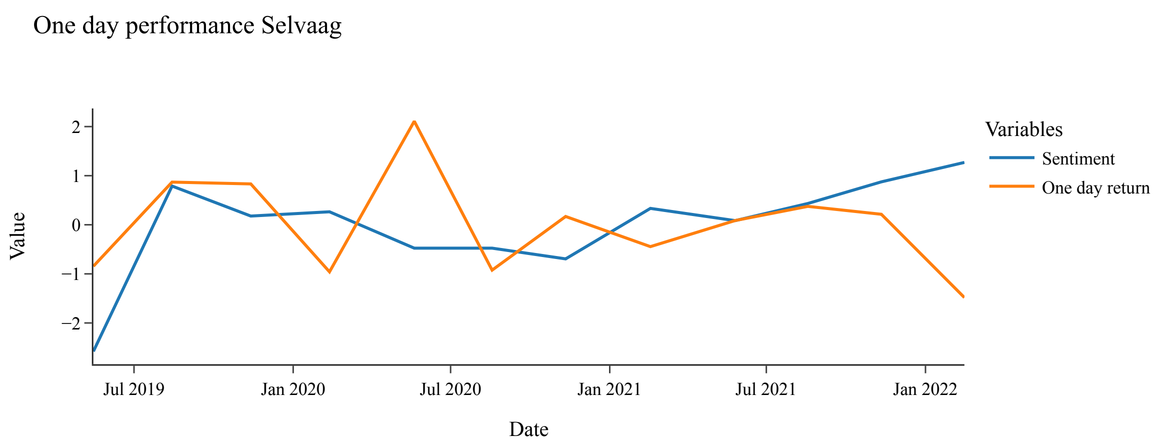


Figure A0.6: One-day performance and sentiment predicted for Selvaag Bolig ASA

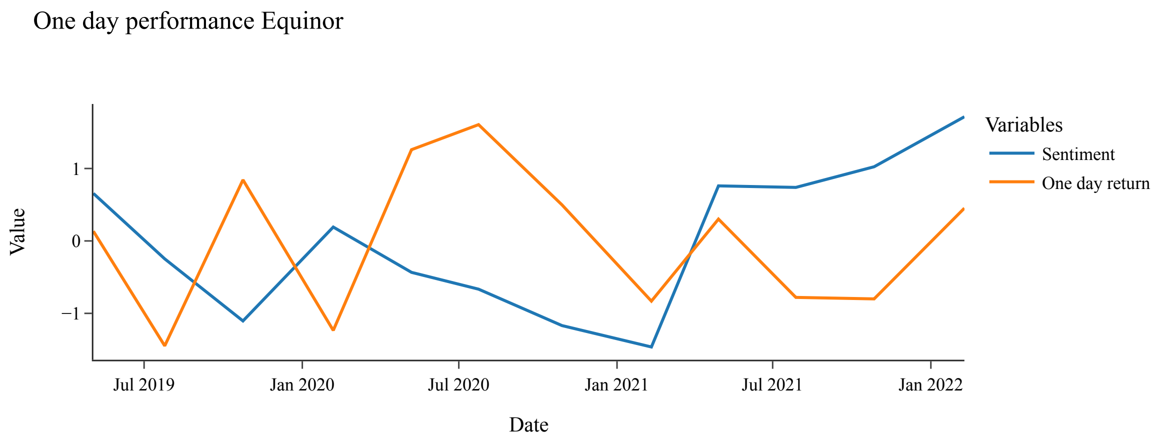


Figure A0.7: One-day performance and sentiment predicted for Equinor ASA

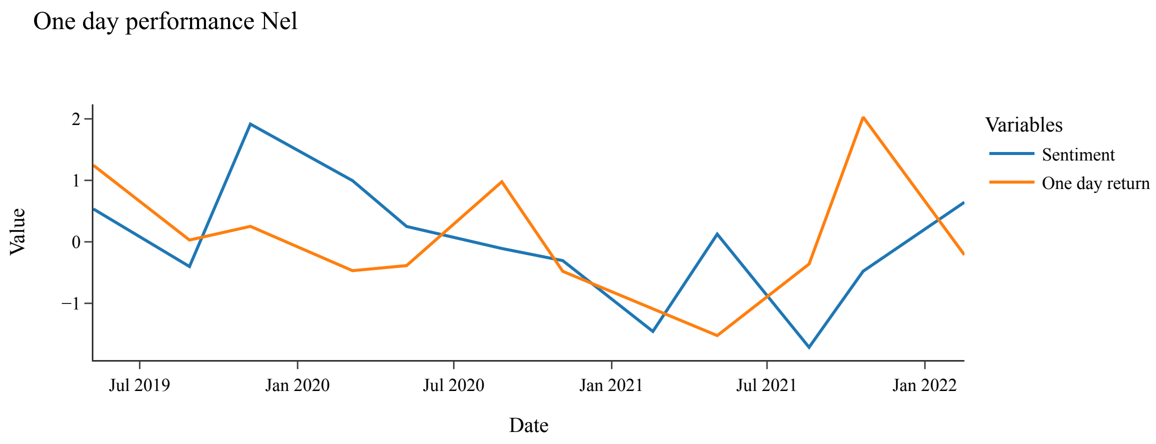


Figure A0.8: One-day performance and sentiment predicted for NEL ASA

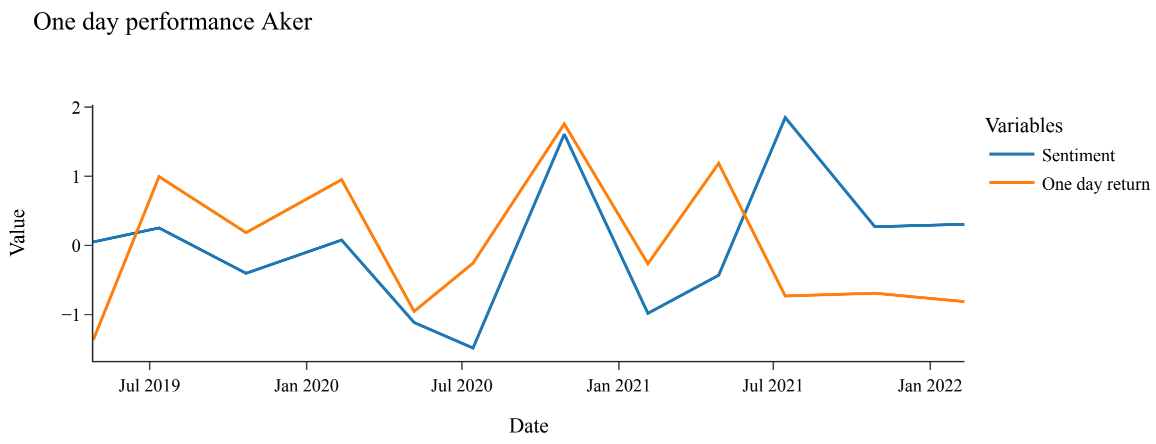


Figure A0.9: One-day performance and sentiment predicted for Aker BP ASA

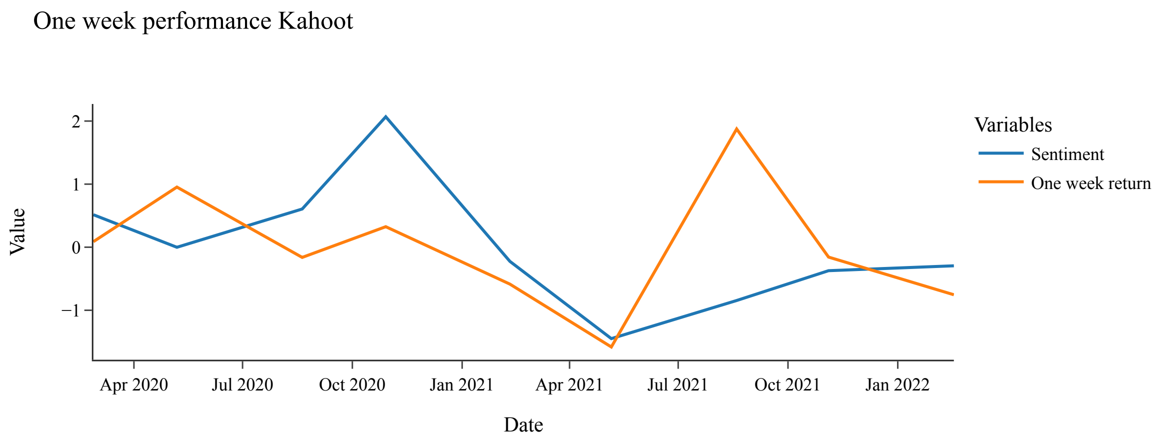


Figure A0.10: One-week performance and sentiment predicted for Kahoot ASA

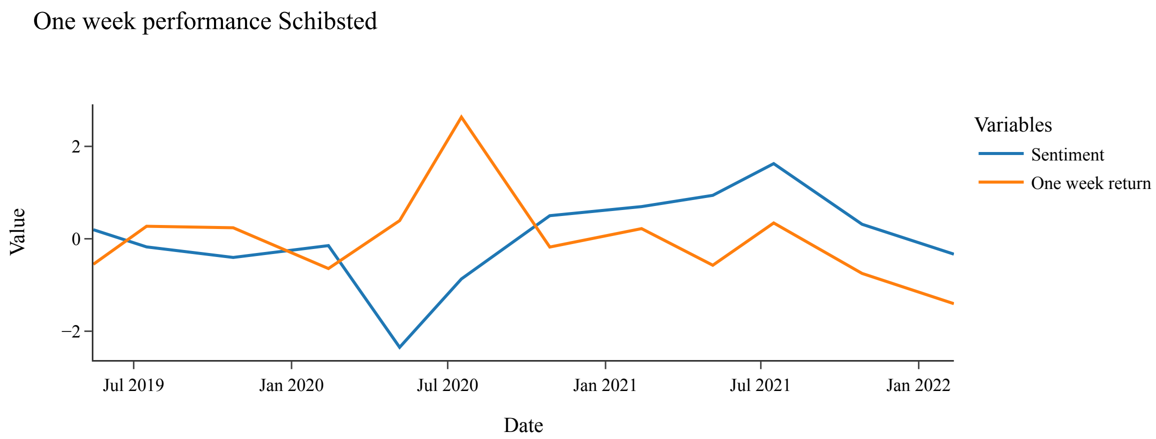


Figure A0.11: One-week performance and sentiment predicted for Schibsted ASA

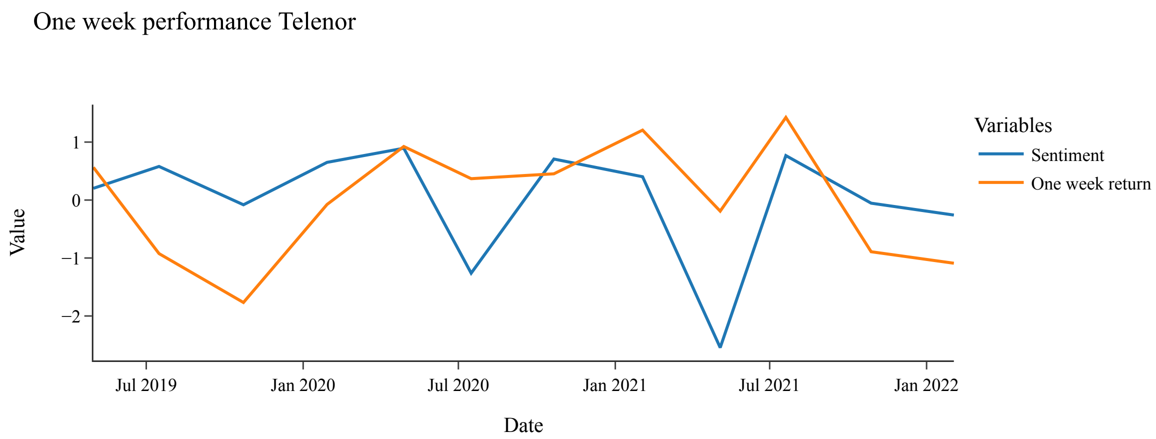


Figure A0.12: One-week performance and sentiment predicted for Telenor ASA

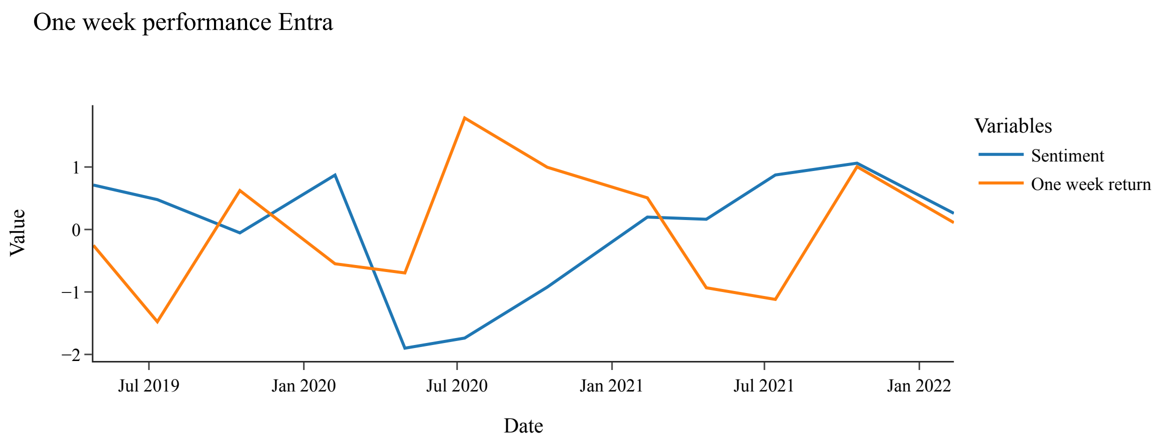


Figure A0.13: One-week performance and sentiment predicted for Entra ASA

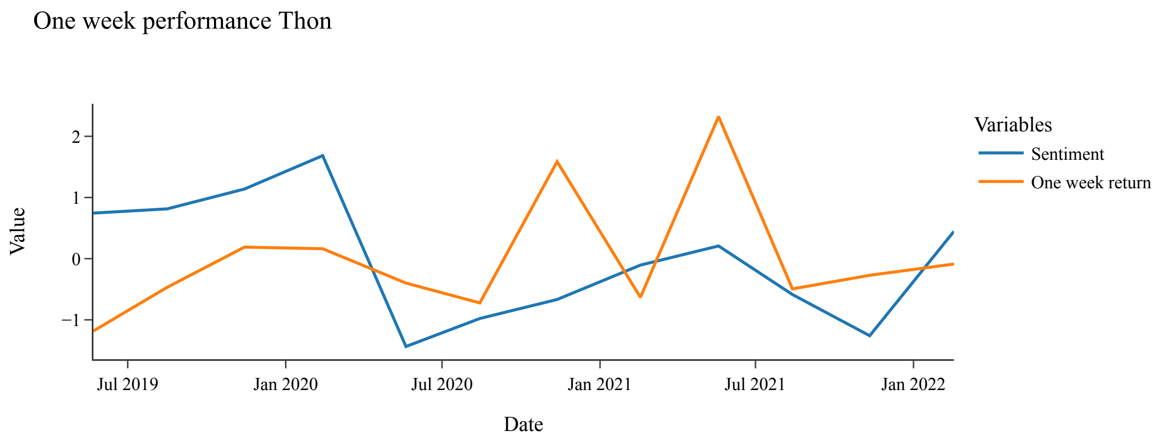


Figure A0.14: One-week performance and sentiment predicted for Olav Thon Eiendomsselskap ASA

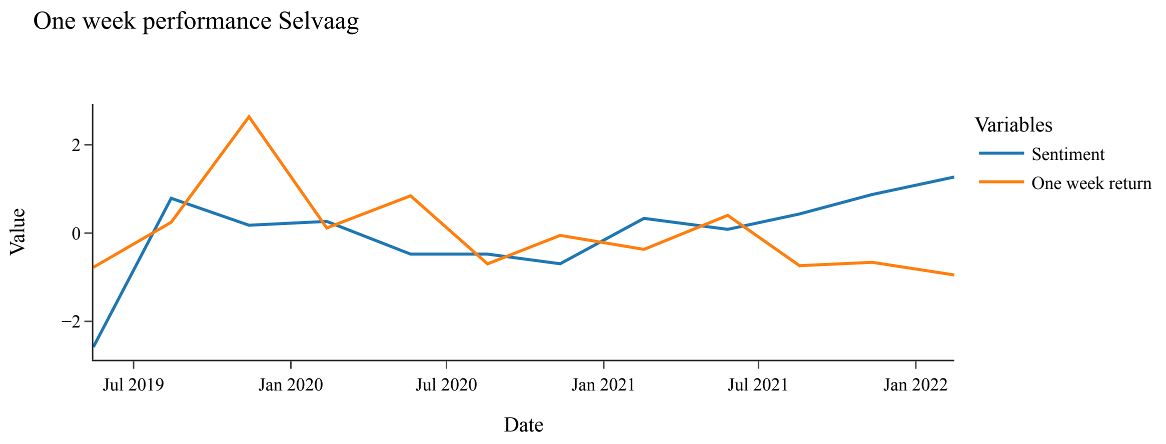


Figure A0.15: One-week performance and sentiment predicted for Selvaag Bolig ASA

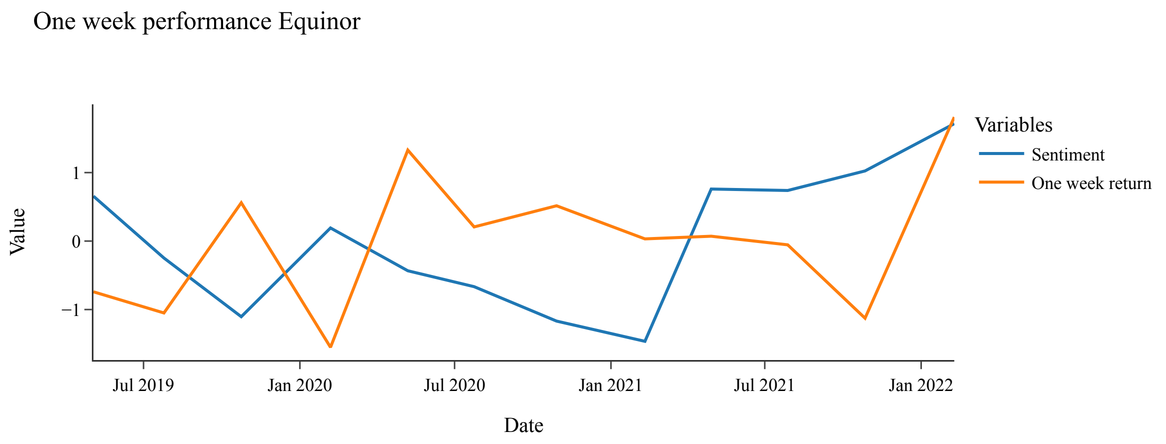


Figure A0.16: One-week performance and sentiment predicted for Equinor ASA

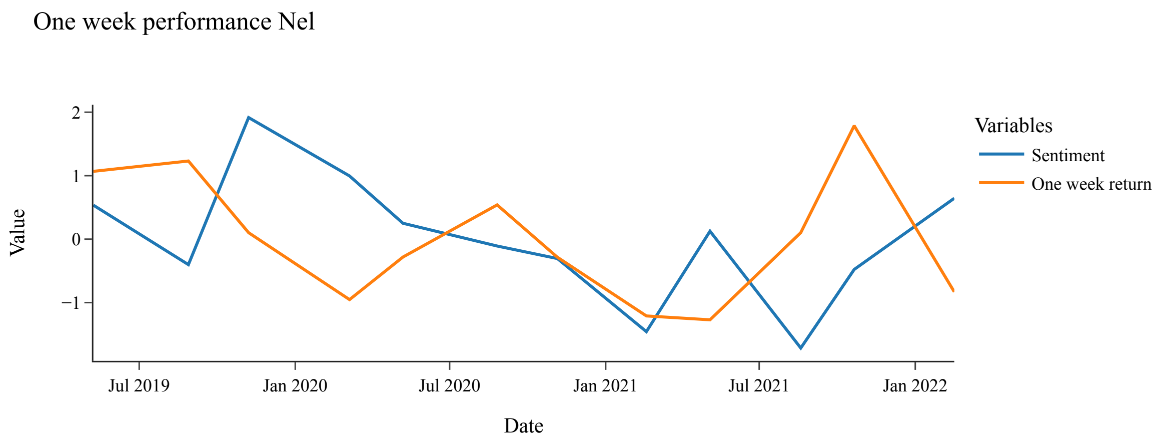


Figure A0.17: One-week performance and sentiment predicted for NEL ASA

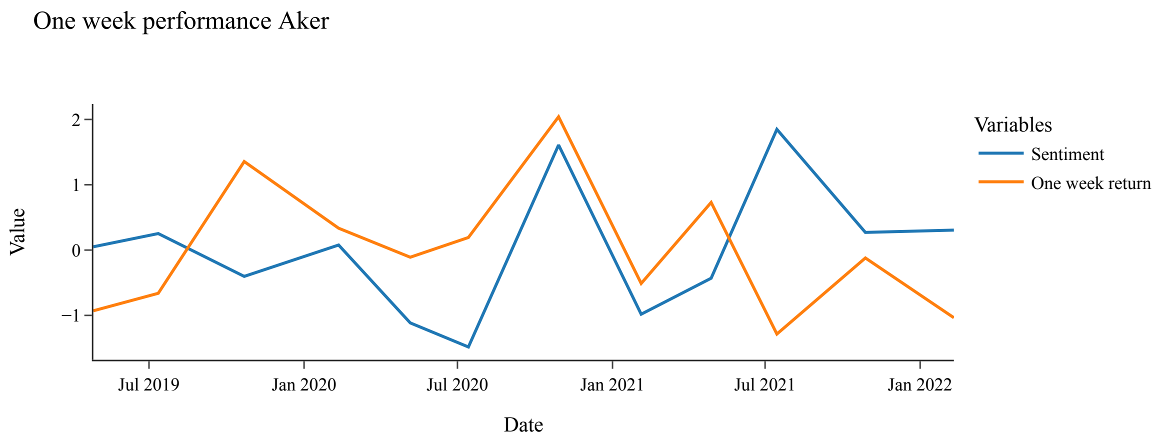


Figure A0.18: One-week performance and sentiment predicted for Aker BP ASA

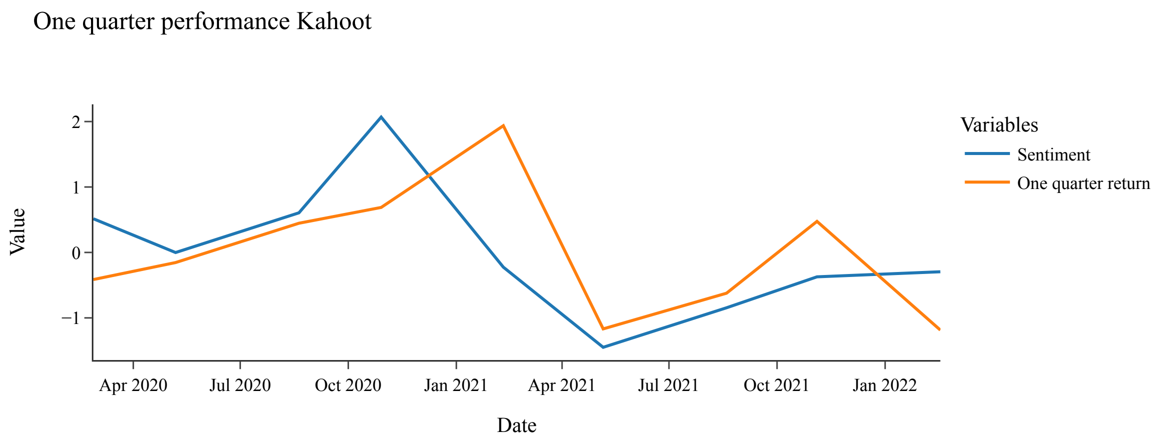


Figure A0.19: One-quarter performance and sentiment predicted for Kahoot ASA

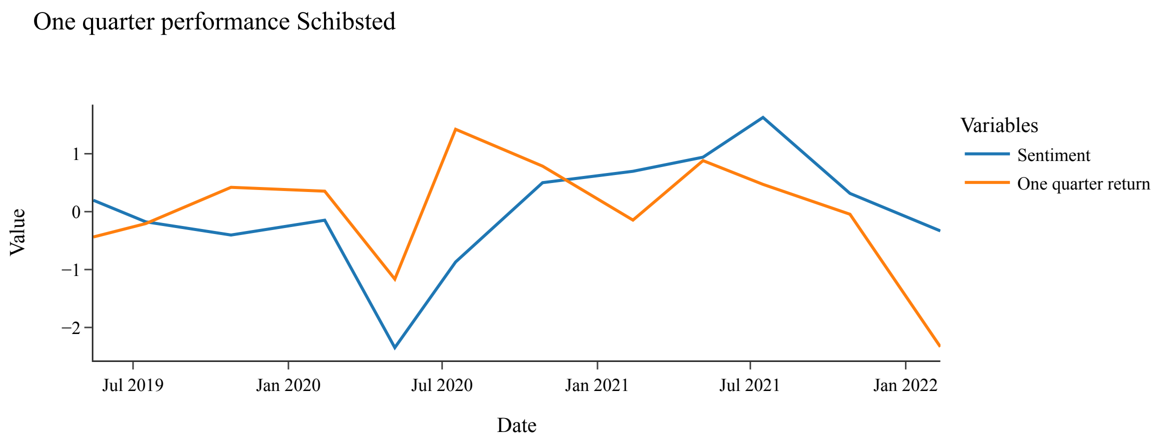


Figure A0.20: One-quarter performance and sentiment predicted for Schibsted ASA

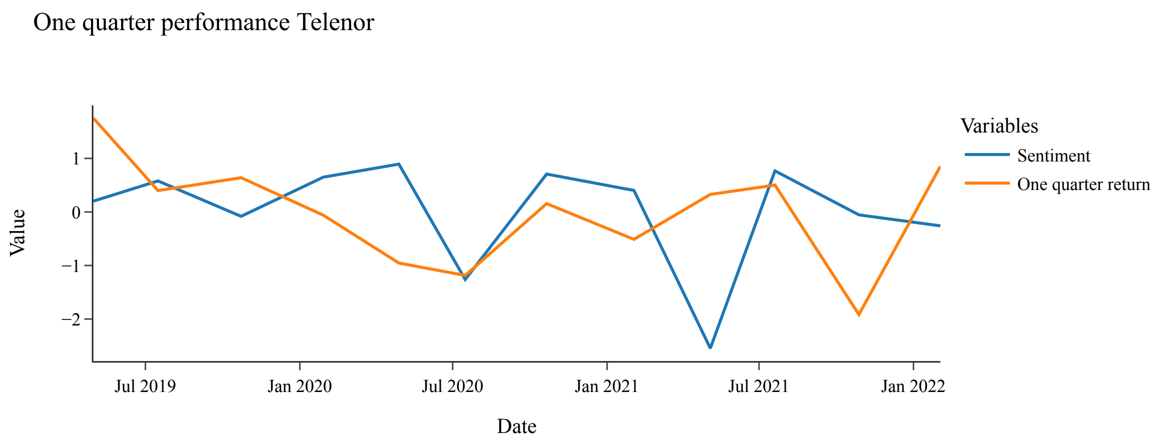


Figure A0.21: One-quarter performance and sentiment predicted for Telenor ASA

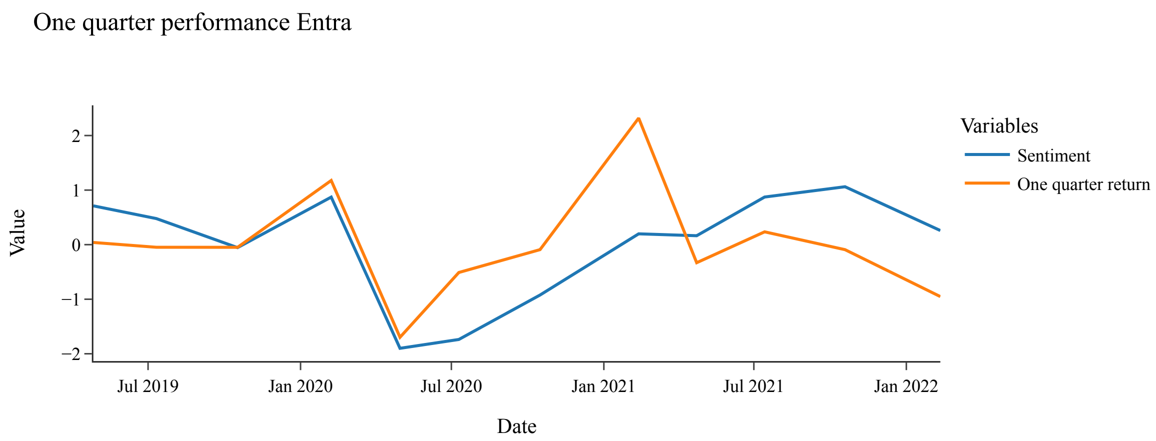


Figure A0.22: One-quarter performance and sentiment predicted for Entra ASA

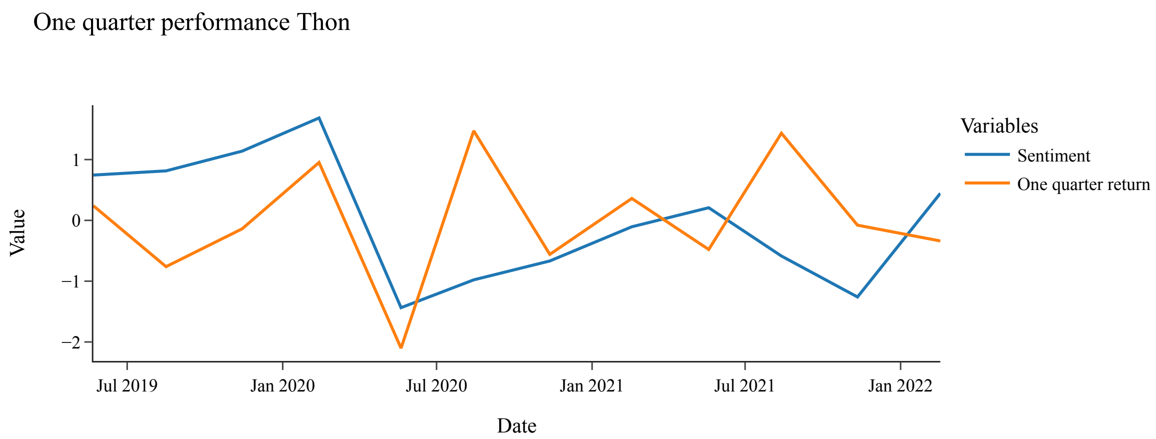


Figure A0.23: One-quarter performance and sentiment predicted for Olav Thon Eiendomsselskap ASA

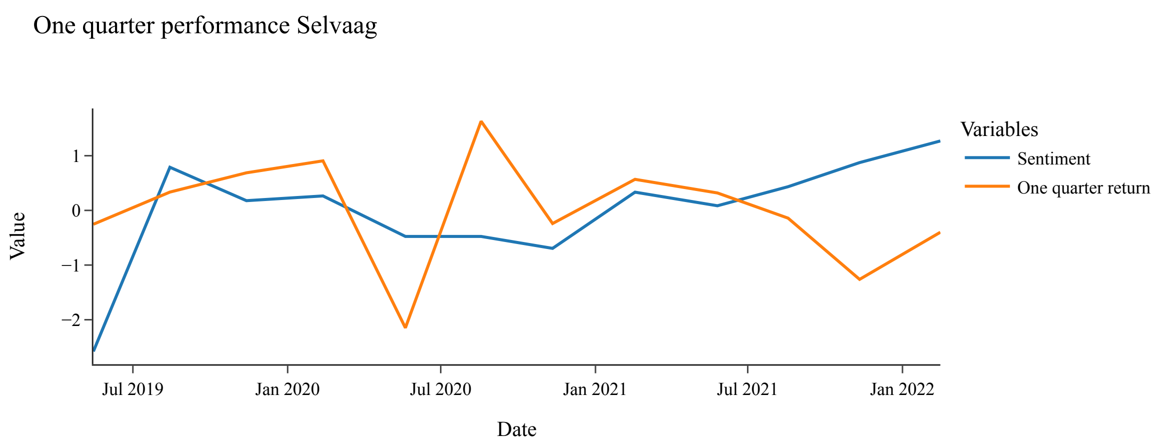


Figure A0.24: One-quarter performance and sentiment predicted for Selvaag Bolig ASA

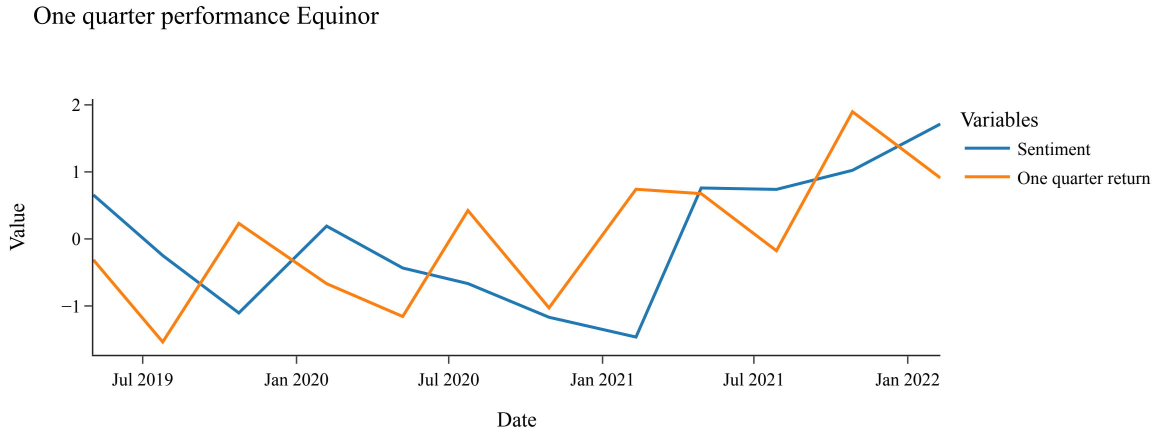


Figure A0.25: One-quarter performance and sentiment predicted for Equinor ASA

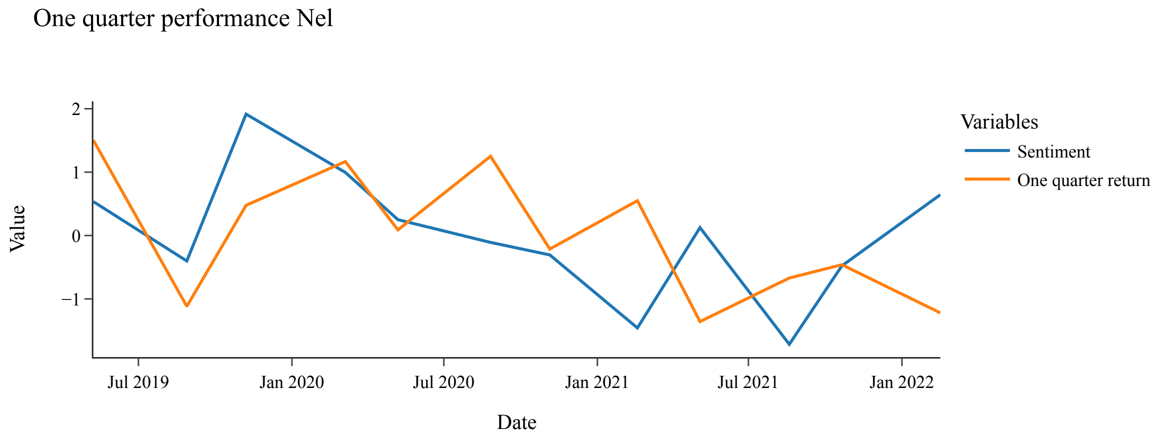


Figure A0.26: One-quarter performance and sentiment predicted for NEL ASA

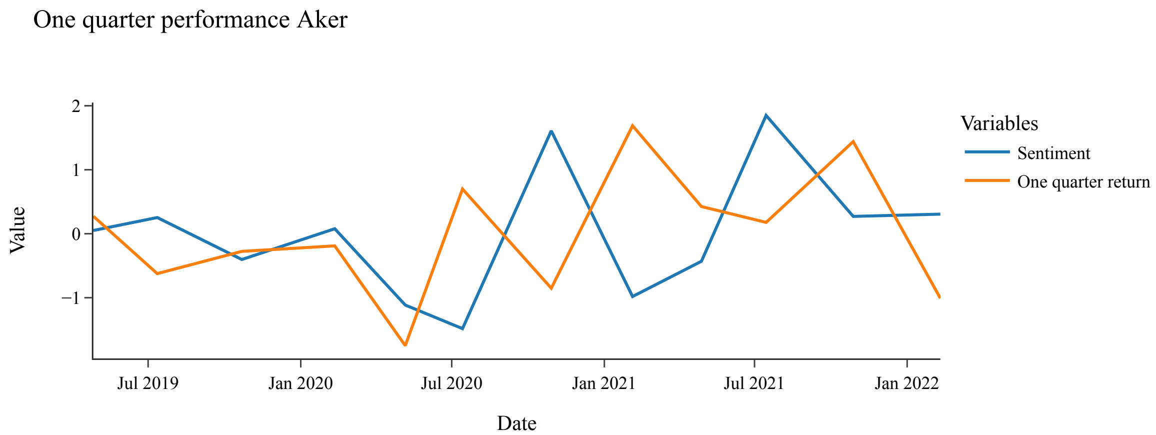


Figure A0.27: One-quarter performance and sentiment predicted for Aker BP ASA

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway