University of Twente

Master Thesis

MSc Business Administration - Financial Management

Effects of Annual Report Sentiment on Stock Returns

Author:
L.A. Borggreve

Student Number:
2632403

Faculty of Behavioural, Management and Social Sciences (BMS)

Dr. D. E. Proksch

Dr. Ir. M. Preziuso

April 19, 2022

UNIVERSITEIT TWENTE.

Acknowledgements

This thesis is written to mark the finalization of my Master of Science in Business Administration with a specialization in Financial Management at the University of Twente. During the writing of this thesis, a lot of people have supported me, who I would like to thank for their support.

First of all, I would like to thank Dr. D. E. Proksch and Dr. Ir. M. Preziuso for their continuous support and feedback during the process. Their valuable contribution to the proceedings and results of this thesis are of such a considerable nature, that it is hard to overstate the impact they have made in compiling this thesis.

Secondly, I would like to thank my colleagues at KPMG. Most of my thesis process has been achieved whilst working with them during the week. During this period, I have always felt their continuous support and understanding for the challenging task of combining writing a thesis with a professional job.

Lastly, I would like to thank my friends and family. As has been the case during any period in my life, they have been a source of energy, motivation, and support, always sticking with me through easier and harder times.

To conclude, I would like to acknowledge the fact that I am very lucky to have surrounded myself with people that have contributed to this thesis process in academic, professional, and social ways.

Leander Borggreve,

Wierden, April 2022

Abstract

This study utilizes natural language processing techniques to analyze annual report narratives. The sentiment of annual report narratives is gauged by utilizing the frequency of words related to a linguistic category to establish sentiment. This sentiment is used to predict abnormal stock returns. We find that the indicators ”positivity”, ”constraining”, and ”superfluous” can, in a model, predict abnormal stock returns. When combined with interaction effects deriving from cultural differences, these models show an even stronger predicting capability for abnormal stock returns. When considering only uni-dimensional models, we find that all of the sentiment indicators have significant effects on abnormal stock returns in the short-term. Among the earlier mentioned indicators, these are: ”readability”, ”uncertainty”, ”litigious”, and ”text density”. These conclusions are drawn after controlling for year- or country-specific trends. Consequently, investors and companies alike are advised to analyze the sentiment of annual reports, as the information included in them is likely to cause short-term adaptations in the stock price.

Keywords— sentiment analysis, stock market returns, annual report narrative

Contents

1 Introduction 5

1.1 Problem background . . . 5

1.2 Research question and relevance . . . 5

1.3 Outline of the study . . . 6

2 Theoretical Framework 7

2.1 Document analysis and its application in finance . . . 7

2.1.1 The evolution of document analysis in finance . . . 7

2.1.2 Recent standards for document analysis in finance . . . 8

2.1.3 Considerations for document analysis . . . 9

2.2 Sentiment analysis . . . 10

2.2.1 Types of sentiment & text . . . 10

2.2.2 Alternative explanatory variables . . . 11

2.2.3 Techniques for analysis . . . 12

2.3 Stock return valuation . . . 13

2.3.1 Cumulative abnormal returns . . . 13

2.3.2 Efficient Market Hypothesis . . . 14

2.3.3 Capital Asset Pricing Model (CAPM) . . . 15

2.3.4 Alternative theories . . . 16

2.4 Hypothesis development . . . 17

2.4.1 Positivity . . . 17

2.4.2 Readability . . . 17

2.4.3 Litigious . . . 17

2.4.4 Uncertainty . . . 18

2.4.5 Constraining . . . 18

2.4.6 Superfluous . . . 18

2.4.7 Text Density . . . 19

3 Method 19

3.1 Sample selection & collection . . . 19

3.2 Data processing . . . 20

3.2.1 File conversion . . . 20

3.2.2 Standardization . . . 20

3.2.3 Tokenization . . . 21

3.2.4 Stemming . . . 22

3.2.5 Word relevance . . . 22

3.3 Variable analysis . . . 23

3.3.1 Independent variables . . . 23

3.3.2 Control variables . . . 25

3.3.3 Dependent variable . . . 26

3.4 Data cleaning . . . 27

3.4.1 Winsorization . . . 27

3.4.2 Normalization . . . 27

3.5 Regression analysis . . . 28

3.6 Supervised machine learning algorithm 29

4 Results 30

4.1 Descriptive statistics 30

4.2 Variable measurement significance 31

4.3 Assumptions 31

4.3.1 Linearity assumption 32

4.3.2 Normality assumption 32

4.3.3 Homogeneity assumption 33

4.3.4 Independence assumption 33

4.4 Correlation 33

4.5 Regression results 34

4.6 Hypothesis testing 37

4.7 Supervised machine learning algorithm 39

5 Discussion and Conclusion 40

5.1 Key findings 40

5.2 Limitations 41

5.3 Future research 42

A Techniques for sentiment analysis 43

B Code listings 43

B.1 PDF to TXT converter (AHK) 43

B.2 Readability analysis (Python) 44

B.3 Dictionary analysis (Python) 44

B.4 Retrieval of actual stock returns (R) 46

B.5 Calculation of expected returns (R) 46

B.6 Optimal number of decision trees for classifier (Python) 46

B.7 Classification algorithm results (Python) 47

C VIF 47

D Confusion matrices 48

References 49

1 Introduction

1.1 Problem background

Everyone is surrounded by some sort of text during most parts of the day, whether you are a tourist reading a guide to explore a new city or a detective reading an interview looking for inconsistencies or clues in the answering of a subject. There is, however, a difference between types of text that people encounter during the day. The guide for the tourist communicates its message as clear as possible, whether the interview read by the detective usually requires a more thorough analysis of the text. These 3 are just examples of the wide array of different types of texts that exist, among all these different types of texts is the annual report.

Annual reports are obligatory public documents issued by corporations for stockholders and other people interested in the company. These reports nowadays offer a plethora of takeaways regarding the state of a company. Information related to, among others, financial figures; potential risks and opportunities; and corporate governance practices are disclosed for the world to see. Historically, the readers of these annual reports, e.g. investors and decision-makers, have shown different approaches to looking at annual reports, which can be divided in two parts: the narrative and the financials. Investors can also choose to look at a combination of the two (Penrose, 2008). These two parts do not always convey the same message, quite the contrary, these two parts have often shown to diverge in worrying ways (Balata & Breton, 2005).

In this research, the focus was placed on the narrative parts of annual reports, which, given their lengthy and sometimes ambiguous nature, have historically been harder to research. However, with the emergence of natural language processing techniques, which can be utilized to analyze these large amounts of text automatically, the sentiment of a narrative can be distinguished more easily. With the sentiment of the text, not merely the positivity is meant, but also subjects like the readability of a text. For example, some annual reports have been described as having low readability, with an estimate of 75% of the U.S. adult population theoretically being unable to comprehend it (Fisher, Garnsey, & Hughes, 2016).

The release of an annual report means a specific moment where investors and decision-makers get hold of information that might influence their opinion on the stock. Therefore, research has been dedicated to finding relationships between different aspects of annual reports and the cumulative abnormal returns of a stock. For example, profitability ratios and market value ratios have shown to have a consistent impact on cumulative abnormal returns (Martani & Khairurizka, 2009).

1.2 Research question and relevance

In order to achieve the goal of this research, which is to define the effects annual report sentiment has on short-term cumulative abnormal stock returns of listed companies, the following research question is formulated:

How does the sentiment of an annual report influence short-term cumulative abnormal stock returns in the US, UK, Germany, and the Netherlands?

The answer to this research question is a valuable contribution to the existing literature by adding a perspective on the different tones and characteristics a narrative can have in an annual report. Much has been written on the sole presence of positivity or negativity in a text, but research has thus far not fully exploited the step of looking at multiple different aspects of text made possible by new technologies and dictionaries. On top of that, most of these prior research efforts have been focusing on one country (Wisniewski & Yekini, 2015; Rahman, 2019), which, given the ease with which a larger dataset is made with the current programming languages, misses out on several potentially important aspects of cultural differences (Aguilera & Crespi-Cladera, 2016). Therefore, it may not come as a surprise that the qualitative information hidden in the narrative is deemed an untapped repository (Fisher et al., 2016). A first look at the variables and data sources to be used in the research can be found in Figure 1.

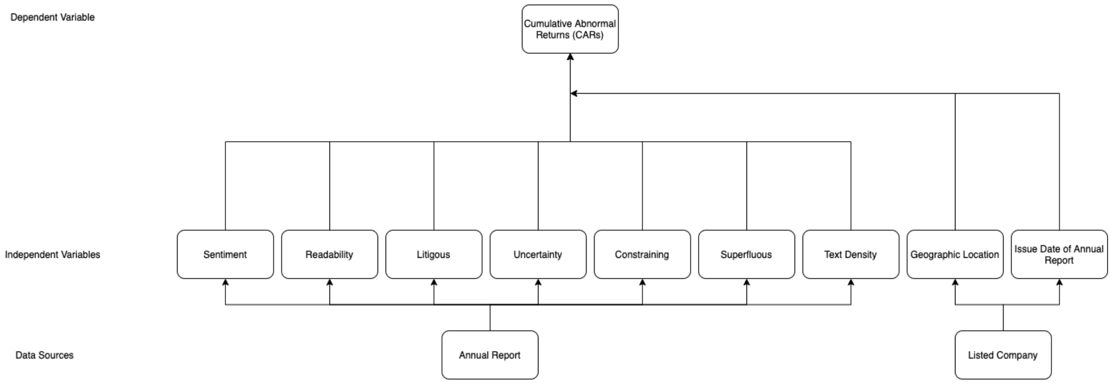


Figure 1: Variables

On top of the theoretical contribution, this research has a practical contribution which can be viewed from 2 sides. First of all, there is a contribution to the way listed companies compose their annual reports. The results derived from this research can give these companies insights into what textual characteristics will yield the best short-term returns for their stock price. Secondly, there is a contribution to the way investors can judge these annual reports. With the help of the prediction algorithm, an investor can utilize the characteristics of a newly issued annual report to quickly calculate the expected positivity of abnormality in stock returns for the short-term more accurately.

1.3 Outline of the study

The remainder of this thesis is structured as follows. Section 2 reviews prior literature and contains the hypothesis development. Section 3 describes the data collection, research method utilized in this thesis, and how the variables are measured. Section 4 contains the analysis of the results. Finally, section 5 presents the conclusion, limitations, and areas for further research.

2 Theoretical Framework

2.1 Document analysis and its application in finance

2.1.1 The evolution of document analysis in finance

The subject of reviewing and evaluating documents is most often referred to as document analysis. The goal is to develop a better understanding of documents or gather knowledge regarding the implications of these documents. Document analysis is far from a new topic, having a history that can be more easily expressed in centuries than in decades, with document usage for text parsing dating back as far as the 1300s (Loughran & McDonald, 2016). As is to be expected, these efforts of document analysis vary heavily from the standards that are to be expected in the current era, with computing power and large corpora aiding those involved in the analysis of documents.

In terms of its application in the financial field, document analysis has known several stages of development. At first, in line with what is to be expected from a technological viewpoint, manual text analyses were conducted in order to analyze financial documents (Fisher et al., 2016). An example of these early research efforts is the research by Poshalian & Crissy (1952), which found that annual reports are ”difficult and dull” from a textual perspective. This type of research stayed the standard until the late 1990s and early 2000s, which is when basic text mining and natural language processing techniques started gaining traction, as can be seen in figure 2. Even though these newer techniques were already available and used in other domains more than a decade before their rise to relevance in financial document analysis, their application in the field was kept to a minimum, with research utilizing NLP like that of Gangolly et al. (1991) being quite the exception. However, from the point that they started gaining traction as a relevant way for analyzing financial documents, these more innovative techniques have continued to develop fast and became the standard in the field (Fisher et al., 2016), with some surveys of techniques relevant to the field disregarding the method of manual text analysis altogether (Kearney & Liu, 2014).

The goals of document analysis in finance as explanatory research have been widespread throughout history and can be classified in several aspect (Loughran & McDonald, 2016). The earliest research efforts were mainly concerned with the prediction of firm performance as a reaction to financial statement disclosures, like Singhvi (1968). Manually, text was analyzed in order to find predictors of firm performance, in line with the earlier observations with regard to document analysis techniques. Some of the variables that were utilized for these earlier predictive studies are still of relevance today. The readability of an annual report, for example, was the cornerstone for predictive research by Soper & Dolphin in 1964, whilst still being used in more recent research (Pajuste, Poriete, & Novickis, 2020). Over the years, both the inputs and outputs of document analysis has changed. In terms of inputs, document analysis has evolved with technology. Documents that have often been used throughout history are documents like EDGAR filings, periodic reports and news releases. Over time, web content and social media updates were added to this list of inputs, their accessibility grew to the point were sample sizes of these inputs reached into the millions (Fisher et al., 2016). In terms of outputs, research stayed quite focused on firm performance predictions, with some other paths that were profiting from the developments in document analysis research.

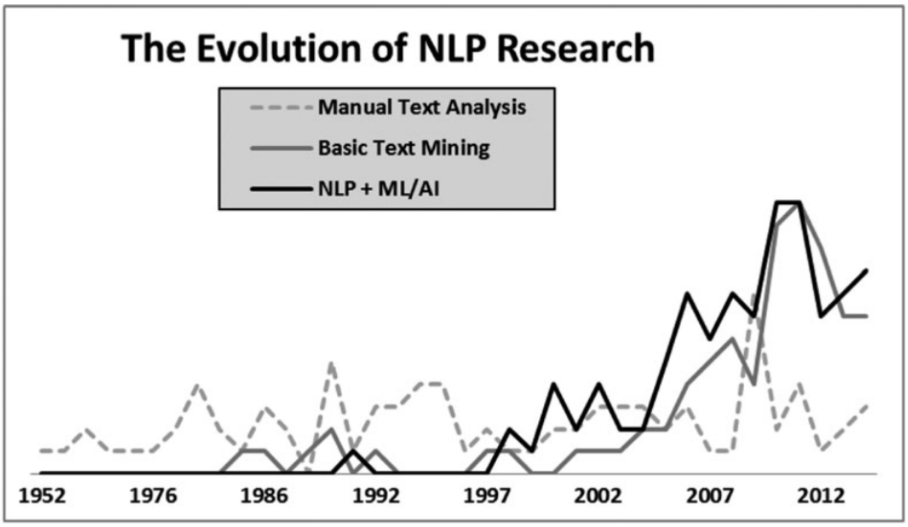


Figure 2: The evolution of document analysis techniques in finance & accounting (Fisher et al., 2016)

One of these paths is the topic of fraud prediction and detection. Starting in the mid 2000s, researchers started applying document analysis techniques in order to predict the risk of fraud; bankruptcy; or other negative events in companies and transactions by the content of the documents involved with these companies and transactions (Woodfield & Irvine, 2006; Santos, Cortez, Pereira, & Quintela, 2006). Another path that has gained in popularity with the rise of document analysis technology, is the prediction of stock prices and other signs of market activity. During the evolution of document analysis in finance, different combinations of explanatory variables and stock prices have been made, often with only limited practical results. Variables often used in this type of research are news sentiment, investor sentiment and, as is the case in this study, annual report characteristics.

Summarized, it is save to say the evolution of document analysis has not only been limited to the techniques utilized, but has also significantly changed with regards to the inputs and predicted outputs of the research.

2.1.2 Recent standards for document analysis in finance

As became apparent from the evolution of document analysis, the more recent standard in the field are text mining and NLP. Before understanding these two techniques, it is important to dive deeper into their different applications. Text mining is the extraction of information from text in an automated way via qualitative analysis (Hearst, 2003), whereas NLP is able of analyzing the meaning of any form of communication, e.g. from a speech (Nadkarni, Ohno-Machado, & Chapman, 2011). From the aspects mentioned in the previous section, the impact of financial texts on financial figures is of importance to consider for this research. For this area of the research, the respected literature has managed to produce low results, when compared with what is normally expected from explanatory research (Falk & Miller, 1992; Chin et al., 1998). These results from research that are trying to explain variances in e.g. abnormal stock returns from document analysis differ quite significantly from source to source. Within the literature research utilizing text deriving from news reports are able of explaining approximately 15-25% of the variance in returns (Wu, Hou, & Lin, 2019; Jeon, McCurdy, & Zhao, 2021), whereas research utilizing text deriving from filings such as annual reports report explained variance statistics varying from a mere 0.1% to relatively more impressive figures like 10% (Wisniewski & Yekini, 2015; Yekini, Wisniewski, & Millo, 2016; Pagliarussi, Aguiar, & Galdi, 2016; Rahman, 2019). As was brought forward by Fisher et al. (2016), the recent challenges to be overcome lie in finding the untapped potential of textual data to optimize investment decision-making. On top of that, it is proposed that the key to tapping into the potential of textual data lies in combining variables that have already been researched, like readability or positivity.

With regards to the historical and more recent trends in the literature, this research should be viewed as a modern and relevant contribution to the application of document analysis to financial documents. It utilizes an input/technique combination that has revolutionized the field over the past decade, deriving from the finance-specific traits of the technique. On top of that, it builds upon the explanatory power of annual report sentiment analysis. This explanatory power, in its current form, has only shown limited impactful results when different types of sentiment were combined, or have not even been researched when looking at cultural impacts. This study seeks to add to the prior literature by adding these relevant and progressive elements to the field.

2.1.3 Considerations for document analysis

The array of documents that has been analyzed in the literature is quite extensive. Examples of these documents are advertisements, books, interviews and, as is the case within this research, annual reports (Bowen, 2009). The analysis of documents, according to Bowen (2009), has several advantages and limitations. The main advantages to be considered when analyzing documents are efficiency, stability, cost-effectiveness, availability, lack of obtrusiveness, exactness, and coverage. On the flip side, the limitations are insufficient detail, low retrievability, and biased selectivity. After these pros and cons are taken into consideration, and their impact on the research is acknowledged, the literature turns their eye to the process of document analysis. According to Karppinen & Moe (2019), the three steps to take for document analysis are (1) research design and identification of document types and sources; (2) accessing, collecting and sampling; and (3) conducting the analysis.

For the identification of document types it is important to first identify what the study aims to achieve, before collecting the documents. In this case, the documents are identified as solely serving the purpose of serving as empirical data for explaining the way the market reacts to annual report issues by listed companies. This, according to the literature, makes the annual reports texts ”worthy of analysis in themselves”. The second identified step of document analysis, the accessing, collecting, and sampling of the data, is concerned with the availability of data. When, as is the case within this research, data is accessible publicly, the issue of accessing is merely a formality to be discussed, but not a challenging hurdle to overcome. The collecting and sampling of the data, however, is a challenge that needs to be discussed more detailed, which will be done in the methodology section. The last aspect to be distinguished in document analysis, is the actual analysis and impact of the research. The analysis of documents should take the characteristics of the documents involved into account, combined with the purpose of the research. Some documents are meant to be read from beginning to the end, whereas other documents are a collection of information that can, to certain extents, be read in any order of chapters. These characteristics are important when designing the research (Karppinen & Moe, 2019), as will become apparent in the methodology.

2.2 Sentiment analysis

2.2.1 Types of sentiment & text

Sentiment analysis in the field of explaining stock return variance can be divided in two types: investor sentiment and textual sentiment (Kearney & Liu, 2014). Investor sentiment is concerned with the subjective beliefs of investors regarding cash flows and risks of investments. This type of sentiment has an established effect on stock prices. Consequently, research is mainly focused on measuring the sentiment and quantifying the effect it has on stock returns (Baker & Wurgler, 2007b; K. Kim, Ryu, & Yang, 2019).

The other type of sentiment, textual sentiment, is focused on the more objective characteristics of text. Often the terms ”sentiment” and ”tone” are used to refer to the positivity of such a text. This, however, does not take into account the fact that sentiment can also include other textual characteristics, like is the case in this research (Kearney & Liu, 2014). For this research, textual sentiment is the most relevant type of sentiment, keeping in mind investor sentiment can have an interfering influence, directly via consumed textual information or indirectly via a ”gut feeling”, on the abnormal stock returns that are observed.

In order to get a better understanding of what textual sentiment entails and in what way it is relevant it is important to get a grasp on the underlying pieces of information. Textual information or sentiment related to finance, according to Kearney & Liu (2014), can be classified into three categories and their respective extraction points: corporation-expressed information that is extracted from public corporate disclosures or filings, media-expressed information that is extracted from media articles, and Internet-expressed information that is extracted from Internet messages. Gandía & Huguet (2021) also include analysts’ reports as a separate information source. However, for this research, the extraction of corporation-expressed information from public corporate disclosures is of importance. From these types of disclosures, like earnings releases and periodical reports, the annual report will be analyzed. In Figure 3 an overview of the types of sentiment and text relevant for this research can be found. The white parts are relevant for this research, the grey parts are outside of the scope of this research.

In the literature, much has been written and researched regarding the sentiment of annual reports. These types of literature focused often solely on the positive or negative tone of the annual report as the independent variable (Yekini et al., 2016) or the change of sentiment between these filings (Feldman, Govindaraj, Livnat, & Segal, 2010). Aside from these efforts, there has been conducted research that takes into account more than just the sentiment of a text alone. However, these research efforts have thus far not made use of the even wider array of textual characteristics that are available for analysis nowadays (Wisniewski & Yekini, 2015; Pagliarussi et al., 2016). More recent research has focused on the improvement of the type of analysis via deep neural networks, taking

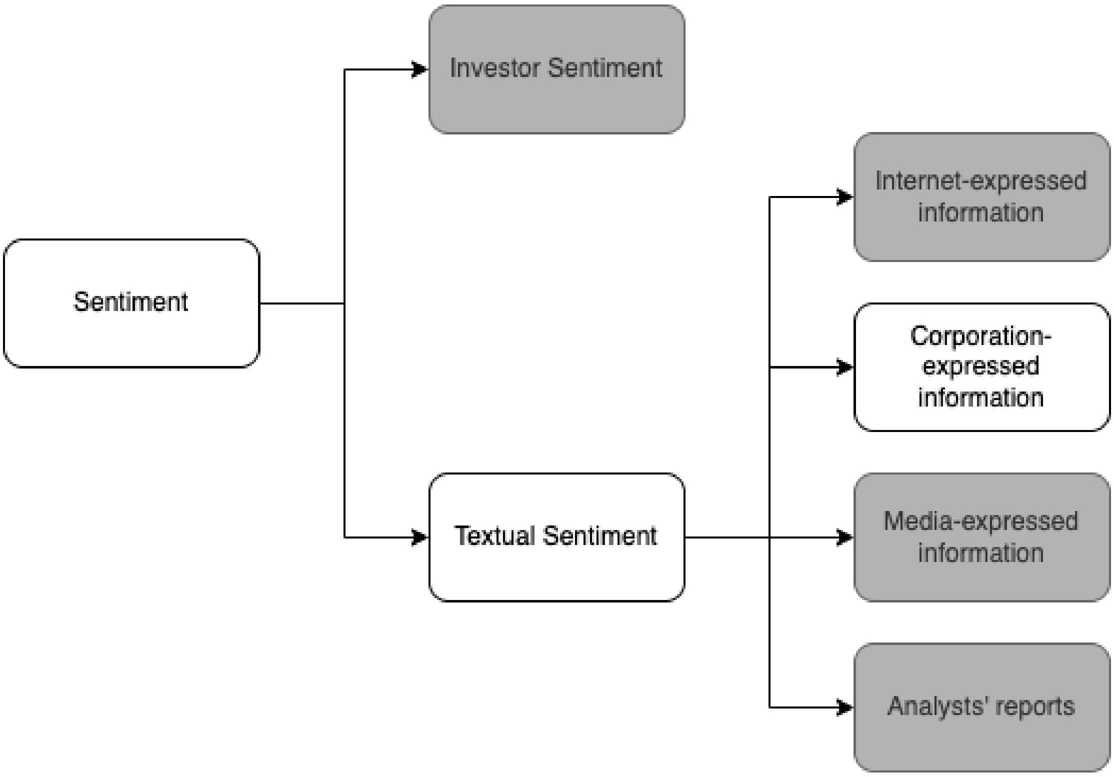

Investor Sentiment
Sentiment
Internet-expressed information
Corporation-expressed information
Textual Sentiment
Media-expressed information
Analysts' reports

Figure 3: Types of sentiment and textual information

the next steps into the progress of applying NLP techniques to potentially predict the effect of the tone of an annual report on cumulative abnormal returns (Wujec, 2021). The research that is conducted is built mainly upon the collective of referenced papers. The papers discussed all researched the same field in one way or another, however, there is a substantial difference in the way the sentiment is analyzed.

2.2.2 Alternative explanatory variables

Since the topic of this research is concerned with annual report sentiment, it is only logical this aspect is thoroughly discussed. However, alternative sentiment variables capable of explaining cumulative abnormal stock returns are of importance as well, in order to properly evaluate the role and place annual report sentiment has in the literature. Since these variables have a prominent place in the literature, it may come as no surprise that the alternative variables sounds familiar to what has already been discussed during earlier sections of the theoretical framework.

Firstly, the topic of investor sentiment can have an explanatory effect on cumulative abnormal returns. Investor sentiment, in short, is the belief investors collectively hold, which is not justified by direct facts or figures (Baker & Wurgler, 2007a). Investor sentiment differs from other types of sentiment in the way that the measurement of this sentiment is quite controversial (Rupande, Muguto, & Muzindutsi, 2019), making it harder to classify it as a clear domain of research with fixed boundaries. However, current research has mainly utilized market-based measures in order to classify investor sentiment, such as; average returns on IPOs, number of new investor accounts, momentum and market skewness (Chakraborty & Subramaniam, 2020; Jiang & Jin, 2021). This type of research is often associated with relatively strong and significant results. However, oftentimes it is also to be debated whether there is causality or only a correlation, since research the other way around is showing similar results (Chakraborty & Subramaniam, 2020).

On top of the financial filing sentiment from sources like annual reports and the investor sentiment, the sentiment of news or social media has a prominent place in the literature. Like mentioned during the discussion of recent standards for document analysis, the analysis of news sentiment in relation to stock returns surrounding the day of the news issue has reaped varying results. These results, in general, have been superior in statistics measurement terms to the analysis of filings like annual reports(Wu et al., 2019; Jeon et al., 2021). Practically, this can be attributed to several factors, of which 2 are the most plausible from a practical viewpoint. First, news reports are more accessible to the public since they are communicated via mainstream media, entering living rooms and offices via all sorts of communication channels. Second, the fact that news reports are consumed quickly after issuance, whereas the majority of accessing of annual reports can happen weeks or months after issuance (Loughran & McDonald, 2017). This means the effect will be more noticeable on the short-term, whereas the effect of annual report issuance can be felt over a longer time period, in line with the time of accessing.

2.2.3 Techniques for analysis

Earlier sections already mentioned that there is more than one technique by which sentiment can be analyzed, in fact, there is quite an extensive array of techniques, which can be found in Appendix A. However, for the analysis of the annual reports, from a literature perspective, two approaches were selected as most relevant: The dictionary-based approach and the machine learning approach.

Dictionary-based approach, A popular approach in the literature is the more traditional approach of using a "word list" or "dictionary" which contains words that are classified as either positive, negative or any other type of classification. It is important to mention that the assumption here is that the words in a text are independent, meaning the order or context of the word is not of any importance (Loughran & McDonald, 2016). This method, often referred to as the "bag-of-words" method, solely relies on the word count of the document to be analyzed. As for the classification of the words, there are several dictionaries available, like the more general and classical General Inquirer (Stone, Dunphy, & Smith, 1966) and DICTION (Sydserff & Weetman, 2002); or the more finance-focused dictionaries by Henry (2008) and Loughran & McDonald (2011). For this research, a distinction has been made between the more general dictionaries and the finance-focused dictionaries. As might sound obvious, prior research has shown that these finance-focused dictionaries show stronger performance when paired with analyzing data that is written in a financial context than the general dictionaries (Henry & Leone, 2010; Loughran & McDonald, 2011; Price, Doran, Peterson, & Bliss, 2012). Between those two more finance-focused dictionaries, the Loughran & McDonald dictionary, by more recent studies, is the more predominant one (Kearney & Liu, 2014), featuring in studies by Doran et al. (2010), Jegadeesh & Wu (2012), and Ferguson et al. (2015).

Machine learning approach, The second relevant approach in the literature is the machine learning approach. Introduced in 2002 as a technique for sentiment analysis (Pang, Lee, & Vaithyanathan, 2002), machine learning as an approach entails a great number of possible sub-techniques. However, these sub-techniques have a lot of similarities in terms of reasons to pick them over a dictionary-based approach, which is why the need to go into great detail regarding the application of these sub-techniques is, for this research, out of scope. In the days before finance-focused dictionaries, the reasoning for application of a machine learning approach was fueled by this absence, since the weakness of using general dictionaries for financial texts was already established (Li, 2010). This reason for choosing a machine learning approach over a dictionary-based approach has decreased over the years. On top of that, the costliness of creating a manually labeled training- and testing dataset makes it, in the current era of big data, a relatively time-consuming task for a thesis. When these datasets are available, the machine learning algorithm is capable of distinguishing the sentiment of the potentially presented texts with relative ease (Che, Zhu, & Li, 2020). This characteristic of sentiment analysis would make it too cumbersome for usage in the actual primary analysis of the data for finding significant effects of the independent variables on the dependent variable. In combination with the dictionary-based approach, however, the application of a machine learning algorithm could prove to be a perfect match for the processed and labeled documents by said dictionary-based approach. Further review regarding the type of algorithm will be discussed in section 3.6.

Several papers have also gone even further, implementing approaches that, in 2019, were classified as the "New approaches" (Shi, Zhu, Li, Guo, & Zheng, 2019). These papers, however, were mostly focused on the classification purpose of the algorithm than on the effects between the independent and dependent variables (Aydogdu, Saraoglu, & Louton, 2019; Wujec, 2021). On top of that, like with the machine learning classification approach, the feasibility of applying these newer techniques on a wide variety of variables within a limited time frame is questionable.

Summarized, the research to be conducted will mainly utilize a dictionary-based approach for the sentiment analysis. However, the machine learning approach will be of relevance for the second part of the analysis, namely the creation of a prediction algorithm, that will be elaborated on further in section 3.6. An illustrated overview of the two described approaches to sentiment analysis can be found in Figure 4

2.3 Stock return valuation

2.3.1 Cumulative abnormal returns

Before reviewing the literature regarding the possible methods for calculating the effect of sentiment on stock returns, it is important to review the literature regarding the measure of adverse stock movements. Whether it is a quick or thorough look at the literature, it is clear the measure of choice for measuring adverse effects on stock movements are (cumulative) abnormal returns (Khin, Tee, & Ying, 2011; Suwanna, 2012; Pagliarussi et al., 2016; K. Kim et al., 2019). This technique calculates how much the actual return of a stock differs from the expected return of a stock for a given time period, hence the "abnormal return". The time periods for these cumulative abnormal returns differ between research. These differences can be attributed to several factors like the relevance

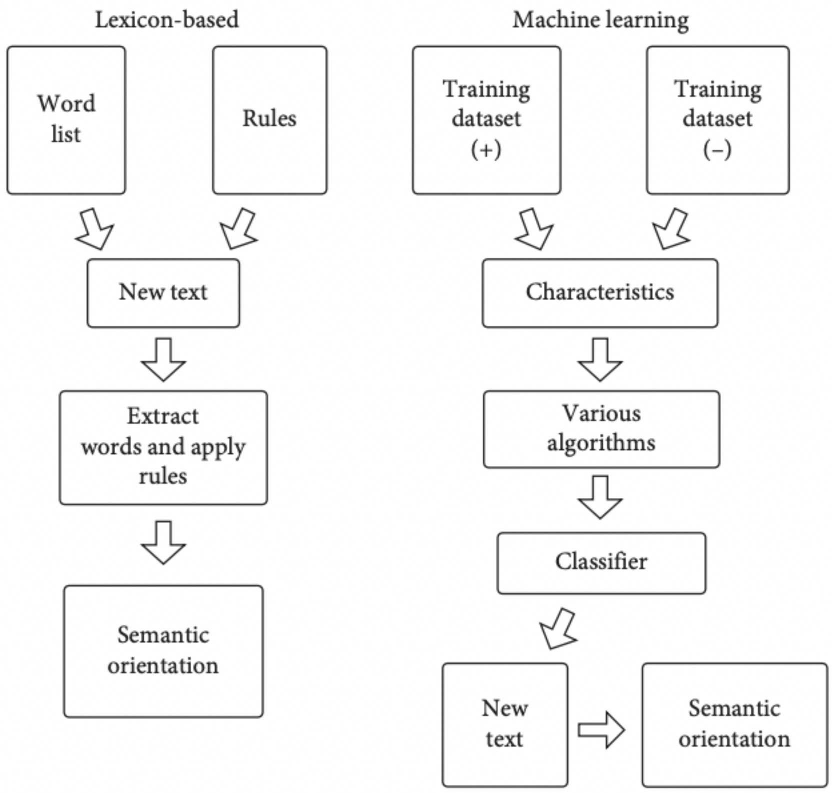


Figure 4: Approach to sentiment analysis (Che et al., 2020)

of post-announcement periods and the short-/long-term focus of the research. As is standard, with the number of days, the number of live trading days is meant. In line with Wisniewski & Yekini(2015) and Kim et al.(2019), a 3-day time period is chosen for this research from a literature standpoint. Furthermore, the more practically substantiated time periods of 5-days and 10-days are chosen, representing the duration of 1 and 2 business weeks, respectively. These longer time intervals are supported by the fact that, in the US, of all the requests in the 400-day period after annual report issue, only 9% of requests are in the first day, 17% are in the first week, and 32% are in the first month after issuance (Loughran & McDonald, 2017). These figures support the notion that a longer time period than the first few days is preferable when researching the effect of annual report issues. During the remainder of this research, the terms ”cumulative abnormal returns” and ”CARs” can be used interchangeably for efficiency reasons.

2.3.2 Efficient Market Hypothesis

The researched effect of annual report sentiment on stock returns is, like almost every research effort, based on several assumptions. An important assumption to discuss is the efficient market hypothesis, which is often referred to as the random walk theory (Kendall & Hill, 1953). This hypothesis is concerned with the degree to which information that is public to market participants is reflected in the current share price (Malkiel, 1989). Since the efficient market hypothesis was theorized in the 60's, there have been 3 degrees of the hypothesis, representing increasing implementations of public and private information in stock prices. In general, it is safe to say that these three degrees entail, from weakest to strongest: only past prices, public information, private information, more clearly illustrated in Figure 5 (Malkiel, 1989; Ying, Yousaf, Akhtar, Rasheed, et al., 2019). For this research, the reflection of public information, like an annual report, into the current share price is an important assumption, otherwise the sentiment of the annual report would not be reflected in the share price.

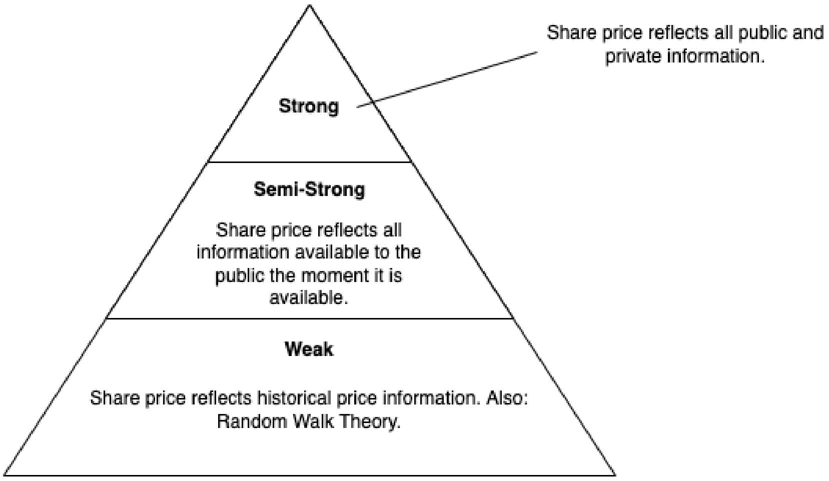
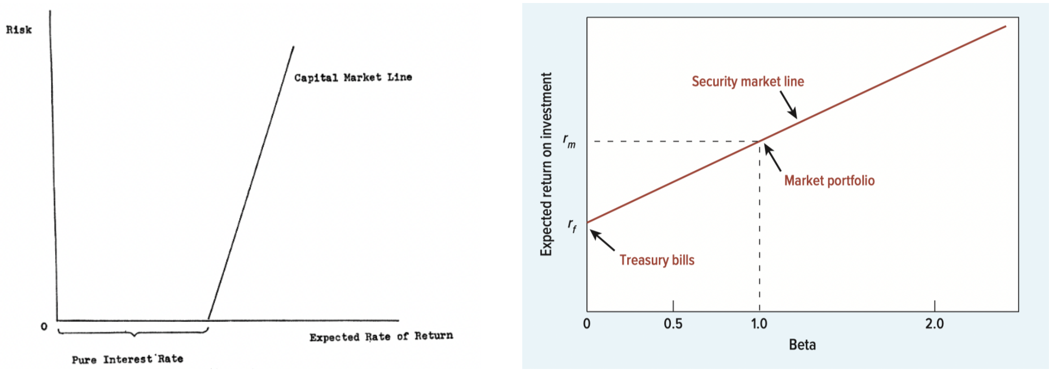

Share price reflects all public and private information.
Strong
Semi-Strong
Share price reflects all information available to the public the moment it is available.
Weak
Share price reflects historical price information. Also: Random Walk Theory.

Figure 5: Degrees of efficient market hypothesis

2.3.3 Capital Asset Pricing Model (CAPM)

In order to properly valuate the independent variable that is the abnormality of returns, a method is chosen and substantiated to calculate this. In the literature, three models are predominantly used: CAPM, Fama-French 3-factor model, and the Fama-French 5-factor model (Pagliarussi et al., 2016; Duan, Zhang, Ding, Chang, & Liu, 2018; Qu, Liu, & He, 2019; Iquiapaza, Carneiro, Amaral, & Ferreira, 2021). From these methods, the CAPM-method is the most suitable for this research. This derives from the fact that the necessity of this research is a method that, on a level, is able to calculate the expected return of a historical period of time for a stock deriving from a historic parameter, which is the β. The CAPM model, first theorized by William F. Sharpe in 1964, is first and foremost about quantifying the return expected when investors take on risk (Sharpe, 1964). The fact that investors require a higher expected return when taking on more risk was known and often represented as the capital market line, which can be seen in Figure 6a. When the capital asset pricing model is discussed nowadays, one is more likely to come across Figure 6b, with a security market line showing the expected return for a given risk/β; a risk-free rate that is often derived from (US) treasury bills; and a market portfolio that represents "the market" and has a β of exactly 1.0 (Brealey, Myers, Allen, & Mohanty, 2018). According to, among others, Pagliarussi et al. (2016), the β can be used as a measure to calculate the expected return of an asset with a certain β as applied in equation 1.



(a) Early (Sharpe, 1964)

(b) Modern (Brealey et al., 2018)

Figure 6: An early and modern illustration of expected return vs risk

$$E(r_{it}) = r_f + \beta_i[E(r_{mt}) - r_f] \tag{1}$$

Where $E(r_{it})$ is the expected rate of return, r_f is the risk-free rate, β_i is the volatility of stock i relative to the market, and $E(r_{mt})$ is the expected return of the market.

2.3.4 Alternative theories

Aside from the CAPM, there are other theories for calculating expected returns. In this section, the attention will be shifted to these other theories and their dismissal for this research will be elaborated upon. The first techniques that will be discussed are the Fama-French 3-factor and 5-factor model. These models are additions on the CAPM model, with the 3-factor adding ”size” and ”book-to-market” to the CAPM model; and the 5-factor model adding 2 more variables to the model: profitability and asset growth (Yang, Li, Zhu, & Mizrach, 2017). By doing this, Fama & French have increased the complexity of the model, whereas the improvement of the model by adding these factors is not unequivocal (Rehman & Baloch, 2016; Coşkun, Selcuk-Kestel, & Yilmaz, 2017). Taking into account the fact that the improvement of results is to be debated and the complexity of the model would be (unnecessarily) increased, the Fama-French models are not used in favor of the CAPM model.

The second theory which is a regular appearance in the literature is the arbitrage pricing theory. This theory takes into account multiple macroeconomic factors and the susceptibility of an asset to these factors to explain the expected return of an asset (Hamao, 1988). The APT model, while in general generating stronger results than the CAPM model (Dhankar & Singh, 2005; Kisman & Restiyanita, 2015), is of such a complexity that only the factor analysis of testing the APT is a research objective on itself (Cagnetti, 2002). For this reason, taking into account the number of countries and years involved in the research, the APT model was dismissed as a feasible measure of expected stock returns.

2.4 Hypothesis development

To really dive deeper into the relevant literature it is important to individually review the variables that are researched by means of sentiment analysis. The chosen variables have been picked for 2 dominant reasons. First of all, in the literature these variables have not sufficiently been researched individually or together with other relevant variables. Secondly, modern day word lists have allowed for the accurate and reliable analysis of these variables via near-exhaustive dictionaries. The variables that are to be discussed in this section are "positivity", "readability", "litigious", "uncertainty", "constraining", "superfluous", and "text density". This list of variables contains the variables relevant for sentiment analysis and a variable that is a side-product from the annual reports, the "text density".

2.4.1 Positivity

The positivity of a text, sometimes referred to as "optimism", is often divided in positivity and negativity. Positivity has shown to positively influence the CARs of the stock price of the issuing company, with Azimi & Agrawal (2021) also advocating for the significance of negative sentiment as a factor (Yekini et al., 2016; Pagliarussi et al., 2016; Azimi & Agrawal, 2021). It is important to notice that the results can be less consistent, given that Pagliarussi et al. (2016) find it as an insignificant factor in their research. Taking all of this into account, this would still result in the following hypotheses:

H1: There is a positive relationship between positivity in annual reports and effect on CARs.

2.4.2 Readability

The readability of a report can be measured using the Gunning-Fog Index (Li, 2008). This index calculates the level of education required for the reader of a text to understand the text. The readability of a text has been shown to influence multiple variables, such as investment efficiency (Biddle, Hilary, & Verdi, 2009), analyst's earnings forecasts (Lehavy, Li, & Merkley, 2011), and trading behavior by small investors (Miller, 2010).

When taking into account the impact readability might have on stock returns, it is also of importance to consider the effectiveness of issuing complex reports. It has been shown that managers that are actually looking to utilize highly complex language in order to hide adverse information from investors are often successful in doing so (C. Kim, Wang, & Zhang, 2019; Pajuste et al., 2020). This is also indirectly confirmed by research by Li (2008), stating that firms with lower earnings often have annual reports that are more difficult to read. The negative relationship between complexity of text and earnings would suggest the following research question:

H2: There is a positive relationship between readability of annual reports and effect on CARs.

2.4.3 Litigious

Another part of the sentiment of texts that can be measured is "litigious". Litigious language can be defined as language that is concerned with taking legal action, one could think of words like "injunction" or "rescinded". However, this aspect of text has not been subject to a lot of research yet when combined with stock returns, whereas it has been used in relation to areas such as business strategy (Lim, Chalmers, & Hanlon, 2018). In the specific case of CARs, of relevance to this research, there is a significant negative relationship between the ”litigiousness” of a text and the abnormal returns of the stock (Pagliarussi et al., 2016). Because of this, the following hypothesis is formulated:

H3: There is a negative relationship between the degree of litigiousness in a annual reports and effect on CARs.

2.4.4 Uncertainty

Uncertainty can be a major part of the narrative of an annual report. Therefore, it has often been researched as an important characteristic of the narrative of an annual report (Hájek, Olej, & Myskova, 2013; Wisniewski & Yekini, 2015; Hájek, 2018). In general, there has not yet been found a significant effect of uncertainty or certainty on CARs (Wisniewski & Yekini, 2015; Pagliarussi et al., 2016). Therefore, within the field of annual report sentiment no substantiated hypothesis can be formulated. However, within the field of investor sentiment a significant relationship has been found for the relationship between uncertainty and stock returns. This research area suggests that (investor) uncertainty has a negative effect on asset valuations, a field that bears close resemblance to that of stock returns (Ozoguz, 2009). For this reason, the following hypothesis has been formulated for this research:

H4: There is a negative relationship between uncertainty in annual reports and effect on CARs.

2.4.5 Constraining

Another analyzed variable is the amount of constraining language in an annual report, this is language that indicates the company is restricted in some ways. Interestingly, constraining language does not correlate highly with the traditional measures of financial constraint (Bodnaruk, Loughran, & McDonald, 2015). The degree of constraining language that can be extracted from the MD&A of an annual report is reported to positively associate with stock returns (Buehlmaier & Whited, 2015). Therefore, in terms of constraining language, the following hypothesis is formulated for this research:

H5: There is a positive relationship between constraining language in annual reports and effect on CARs.

2.4.6 Superfluous

Superfluous language in documents like annual reports is not a new topic to be discussed, the SEC has formally required firms to utilize ”plain English” in filings from 1998 onward (Loughran & McDonald, 2014). It is important to notice that this does not mean that complex data is omitted, it should merely be presented in a more orderly way (Tavcar, 1998). Empirically, it is true that the SEC policy enacted in 1998 has drastically reduced superfluous language in the majority of annual reports, this effect was enhanced even further when the Sarbanes-Oxley act was enacted in 2002 (Loughran & McDonald, 2014). However, literature until now has been remiss in researching the effect that superfluous language still has on the returns of stocks. Based on the fact that is has become clear that superfluous language is looked down upon, the hypothesis is formulated as follows:

H6: There is a negative relationship between superfluous language in annual reports and effect on CARs.

2.4.7 Text Density

Text density, the amount of words per page, is a variable that bears close resemblance to readability and is often used as a part of readability (Li, 2008). This can be attributed to the fact that text dense reports, with long and complex sentences, have a negative effect on readability. On top of that, this ”impression management” is known as reading ease manipulation (Ajina, Laouiti, & Msolli, 2016). Given the negative sentiment surrounding this practice, in combination with the negative relationship hypothesized for ”readability”, the hypothesis for the text density variable is as follows:

H7: There is a negative relationship between text density in annual reports and effect on CARs.

3 Method

3.1 Sample selection & collection

The research method consists of a combination of researched literature and the traditional steps necessary for textual analysis for most variables. First of all, it is important to retrieve the annual reports and elaborate on their characteristics. The research will be conducted within 4 different geographic areas; US, UK, Germany, and the Netherlands. These 4 countries all have developed markets with a considerable amount of listed companies. From each country approximately 75 listed companies are selected from the major exchanges. Of these companies, where possible, the annual reports from the years 2017-2020 are considered for textual analysis. This would result in a starting sample size of approximately 1000 annual reports to analyze in terms of positivity, readability etc. This number is in line with standards in the literature (Wisniewski & Yekini, 2015; Rahman, 2019). These annual reports will be retrieved from the official website of the respective company. The selection of companies is a delicate process, with the risk of adverse selection, whether intentional or unintentional, present at every decision made regarding research design. In order to minimize the effect of this limitation, a top-down approach is taken with regards to the market cap of companies involved. For larger indices, like the FTSE 100 in the UK, this means taking the largest companies from that index. For countries with smaller indices, like the Netherlands, whose AEX only has 25 constituents, this means working from large-to-small index until the sample is filled. Fortunately, annual reports of listed companies are public information, so the accessibility of them is no issue. This makes the collection of the annual reports as straightforward as downloading them from the company’s website.

In terms of the cumulative abnormal returns, these will be extracted from the database of Yahoo Finance and Refinitiv Eikon. All of the relevant figures: the returns of the stock, the β, and the return of the market of the relevant countries can be found in these databases. These figures are all in some way important to calculating the CARs, which will be elaborated upon in section 3.3.3.

3.2 Data processing

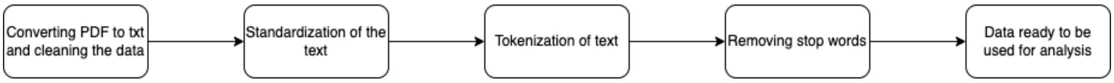

Converting PDF to txt and cleaning the data
Standardization of the text
Tokenization of text
Removing stop words
Data ready to be used for analysis

Figure 7: Pre-processing of the data for analysis

3.2.1 File conversion

These annual reports are often published in PDF format, meaning they need to be converted into a txt file before any analysis can be performed. For this, a tool is created in AutoHotKey[1]. This tool essentially automated the process of manually converting the PDFs to txt files and can be studied in more detail in Appendix B.1. During this process, a small amount of data had to be discarded for reasons deriving from the formatting of the annual report. Annual reports that were containing types or fonts of text that were too difficult to recognize in PDFs resulted in txt files that were too impaired to be usable in this research. Solving or creating these txt files manually would be a task too labor-intensive for this thesis, aside from the fact that the preserved data is still a sizable amount. Data was discarded throughout the process, meaning certain cases of corrupted data were picked up during the converting, whilst others were picked up during outlier analysis of e.g. the variable ”readability”. In table 1 an overview can be found of the preserved and discarded data sorted by geographic location.

Table 1: Preserved and discarded data per country

	USA	United Kingdom	Germany	The Netherlands	Total
Preserved	239	242	216	233	930
Discarded	11	7	44	7	69
Total	250	249	260	240	999

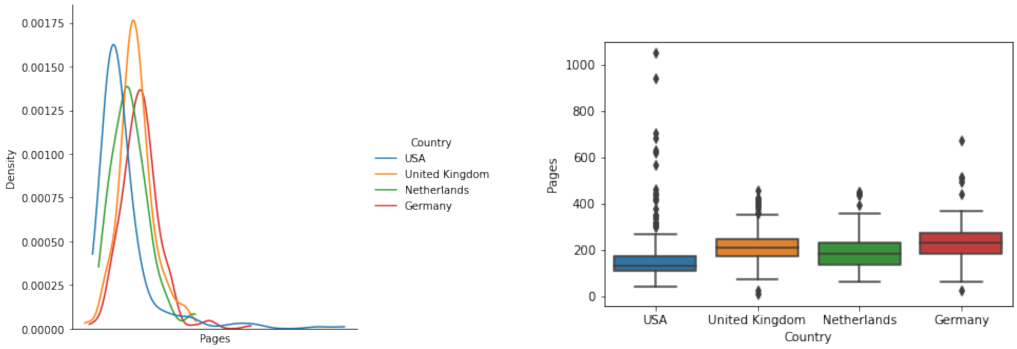
Note: This table represents the preserved and discarded data after file conversion. Subsequent discarding of data on other grounds is not yet included.

3.2.2 Standardization

After the data is converted to text, more steps are required to process the data. These steps are also visualized in Figure 7. The code for these processing steps can be found in appendix B.3. This includes standardizing, or normalizing, the text, which consists of the following steps: making all text lowercase, removing punctuation, removing numbers, and removing whitespaces. This step creates tokenizable txt files that are workable. In the early stages of researching a subject with regards to NLP, simplicity of data and models is important, which is exactly what these steps aim to achieve (Brownlee, 2017). However, before the data was normalized, a first look was taken at the data regarding

[1] AutoHotKey (AHK) is an open-source programming language based on scripts that can automate tasks on Windows, tasks like filling in forms, auto-clicking, and other repetitive assignments.

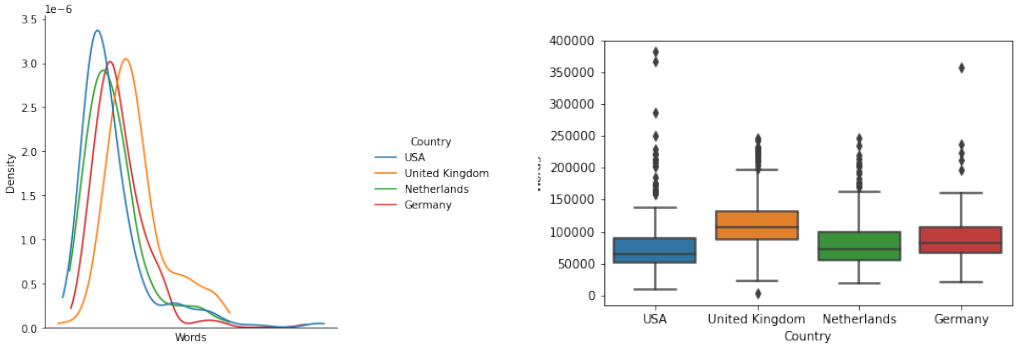
the presence of outliers in terms of the number of words or pages by means of density plots and boxplots, again sorted by country. As can be seen in these figures, outliers are present. Here is acted upon accordingly, as will become apparent later in this section.



(a) Density plot

(b) Boxplot

Figure 8: Number of pages per country



(a) Density plot

(b) Boxplot

Figure 9: Number of words per country

3.2.3 Tokenization

Afterwards, the txt file is tokenized, this means storing each word in the txt file as an independent value in a list. This step is necessary for utilization of the ”bag-of-words” method. The bag of words model requires a piece of text to be split in separate words, also known as tokenization. This process can be done with individual words (1-gram), like ”profitable”, or with word combinations (n-gram), like ”extraordinary profits”. For the scope of this research, the more simple application of the 1-gram technique is sufficient to determine the tone of an annual report. On top of that, when using a dictionary approach, the possibility of using n-grams other than 1-gram requires a dictionary that accounts for these combinations of words, which is not readily available for financial texts.

3.2.4 Stemming

The possible next step is to stem the words, restoring words with suffixes to their original root, like "profitable" to "profit". However, this step is not completely undisputed, with research indicating there are words where stemming changes the meaning of the word, like "odds" turning into "odd" (Loughran & McDonald, 2011). In order to avoid this negative effect deriving from stemming, the finance-specific dictionary of Loughran & McDonald is used, excluding the practice of stemming from this research. After these steps the data is ready to be analyzed. Even though the research does not apply the technique of stemming, it would be remiss if the technique and its application was excluded in the method.

3.2.5 Word relevance

Annual reports, as can be seen in the previous sections, consist of a large number of pages and an even larger number of words. It is in the nature of language that not all words in a sentence or text have great relevance to the message being conveyed. A great number of words is merely utilized to construct sentences that sound correct and organized, making sentences in line with what is expected from a grammatical viewpoint. These words, in sentiment analysis, solely create "noise" around the variables that are actually relevant for the analysis. Therefore, these words should be addressed by one, or a combination, of the below mentioned techniques.

Stop words. Words without relevant information can be removed. Words like "the" and "or" do not have any value in this research and can thus be omitted. These words can also be described as "stop words", which can be filtered out by means of automation, often with Python packages such as "NLTK stop words" or the stop words mentioned in the Loughran & McDonald dictionaries (Müller & Guido, 2016). These packages automatically remove these stop words from tokenized lists.

TF-IDF and Zipf's law. The step of removing stop words is often mutually exclusive with a practice called "Term Frequency - Inverted Document Frequency" or "TF-IDF", which calculates the relevance of words in a document relative to their appearance in the total set of documents (Ramos et al., 2003). For example, a word that appears 100 times as often as another word is not very likely to be 100 times as important as that other word (Loughran & McDonald, 2011; Pagliarussi et al., 2016). This might sound exaggerated, however in Table 2 the frequency of the top 3 used words in the Dutch corpus are shown together with the frequency of 3 random words that can be used as positive or negative words and derive from the same corpus. This table shows a phenomenon known as Zipf's law, which is concerned with the fact that the frequency of word usage follows a "1/rank of the word" distribution, which is also closely related to the Pareto principle (Shyklo, 2017). Also, in Figure 10 is shown the logarithmic distribution of words and their rank in terms of frequency of one particular annual report in the Dutch corpus, once again illustrating the presence of Zipf's law in this study.

Table 2: Frequency of words in Dutch corpus

Word	Frequency
the	1265913
and	688108
for	215135
good	3971
down	3959
improved	3271

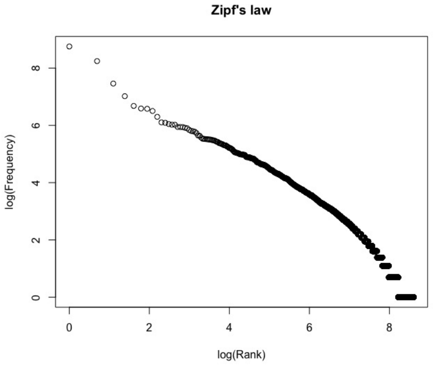


Figure 10: Zipf's law

As mentioned, this effect can be reduced by applying TF-IDF or "term weighting" to account for word frequency. The formula that can be used to achieve this is as follows:

$$\begin{cases} W_{i,j} = \frac{(1+log(tf_{i,j}))}{(1+log(a_j))} log\frac{N}{df_i} & \text{if } tf_{i,j} \geq 1 \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

3.3 Variable analysis

3.3.1 Independent variables

Now that the data is processed it is ready to be analyzed on the independent variables identified earlier. For this, there are several packages, methods, and dictionaries used in accordance with the literature. Between the variables, a difference needs to be made between the textual variable of "readability" and the other analyzed sentiment variables, this can also be seen in Table 3.

For the variables POS, LIT, UNC, CON, and SUP a lexicon or dictionary is used. For this research, it is not necessary to build the dictionary manually, since the chosen approach includes utilization of the pre-set dictionary as composed by Loughran & McDonald (2011), which is constantly being updated. The code for this process can be found in Appendix B.3. The amount of words that the dictionary entails for a specific variable can be found in Table 3. Two variables of the dictionary were dismissed for research given their low number of entries in the dictionary, these were strong-modal and weak-modal, with 19 and 27 entries, respectively. As mentioned earlier, utilization of a finance-oriented dictionary, as opposed to regular word dictionaries, reduces the chance of ambiguity in the meaning of words. The word "cold" would be positive when talking about a fridge, but negative when talking about a coffee machine, to give a general example of possible ambiguity. In a financial context, word like these could be "depreciation" and "tax". The calculation of the variable POS, while straightforward like the others, differs from the other Loughran & McDonald variables. This can be explained by the fact that the POS variable includes usage of 2 dictionaries, positive and negative words, whereas the other variables only utilize 1. This results in the calculation for positivity, as shown in Equation 3.

Table 3: Independent variables and the methods of analysis

Variable	Abbreviation	Explanation
Positivity	POS	The variable ”Positivity” is concerned with measuring the positivity in an annual report via the Loughran & McDonald dictionary. It utilizes a positive (354 words) and negative (2355 words) dictionary.
Readability	REA	The variable ”Readability is concerned with measuring the readability in an annual report via the Gunning-Fog-Index.
Litigious	LIT	The variable ”Litigous” is concerned with measuring the amount of litigious text in an annual report via the Loughran & McDonald dictionary (905 words).
Uncertainty	UNC	The variable ”Uncertainty” is concerned with measuring the uncertainty in an annual report via the Loughran & McDonald dictionary (297 words).
Constraining	CON	The variable ”Constraining” is concerned with measuring the amount of constraining text in an annual report via the Loughran & McDonald dictionary (184 words).
Superfluous	SUP	The variable ”Superfluous” is concerned with measuring the superfluousness in an annual report via the Loughran & McDonald dictionary (56 words).
Text Density	TEX	The variable ”Text Density” is concerned with measuring how many words on average are on a page in an annual report.
Culture	CUL	The variable ”Culture” is concerned with establishing the culture in the country of listing for a stock.

$$\text{POS} = \frac{\text{Number of positive words - number of negative words}}{\text{Total number of words}} \tag{3}$$

For the variables utilizing only 1 dictionary, which are: LIT, UNC, CON, and SUP, the calculation is as follows:

$$\text{Independent variable} = \frac{\text{Number of words of the variable class in the document}}{\text{Total words}} \tag{4}$$

For the readability (REA) of the annual report, the Gunning-Fog-Index is used (Li, 2008; Hájek, 2018; Bai, Dong, & Hu, 2019). The Gunning-Fog-Index is a value that ranks the readability of a text on a scale of ”reading level”, often made more concrete by applying it to the levels of the US education system. For example, a value of 7 means the level of education of someone in the seventh grade is likely necessary to understand the text, whereas a value of 13 means the text is probably only comprehensible by someone that is at least a ”college freshman” in the US education system. The calculation of the

Gunning-Fog-Index is automated in Python via usage of the textstat package, for which the code can be found in Appendix B.2. The equation used for the Gunning-Fog-Index, and therefore in this Python package, is as follows:

$$\text{Fog-Index} = 0.4 * \left[\left(\frac{\text{words}}{\text{sentences}} \right) + 100 * \left(\frac{\text{complex words}}{\text{words}} \right) \right] \tag{5}$$

There are also independent variables that do not require such extensive pre-processing of data: the text density and the culture. The text density variable is calculated as the average amount of words per page, as illustrated in the following equation:

$$\text{Text Density} = \frac{\text{Total number of words}}{\text{Total number of pages}} \tag{6}$$

The last variable is culture. The cultural value of an observation is directly linked to the culture in the country where the stock is listed. In line with Aguilera & Crespi-Cladera (2016), the distinction is made between Anglo-Saxon culture (US & UK) and Continental European culture (Germany & the Netherlands). This variable was mainly utilized as an interaction variable further explaining the differences not only between stock returns, but also between the different cultures.

3.3.2 Control variables

In order to properly evaluate the relationship between the dependent and independent variables, it is important to control certain variables that might impede with the reliability of the analysis. One of these variables is the presence of systematic risk, risk deriving from general movements in the markets as a result of macro-economic events or other wider events triggering investor actions en masse. In order to control for this systematic risk, it is included as a deductible in the calculation of the abnormal returns. This deduction is equal to the movements of the entire market on that particular day. When the entire market is mentioned, the market of the country in which the company is primarily listed is meant. This distinction still accounts for worldwide events, whilst also taking into account country-specific events like elections, national holidays or nation-wide scandals. The difference between these control variables embedded in the abnormal returns, separated by country, can be found in table 4.

Table 4: Summary statistics of market return per country

Country	N	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
Germany	1262	0.000323	0.0121	-0.122	-0.00438	0.000679	0.00581	0.110
Netherlands	1279	0.000443	0.0105	-0.108	-0.00400	0.000851	0.00546	0.090
United Kingdom	1264	0.000098	0.0115	-0.120	-0.00446	0.000382	0.00558	0.103
US	1258	0.000995	0.0148	-0.128	-0.00465	0.00138	0.00801	0.108

In order to reduce the effects external variables can have on the results of the research even more, an extra control variable is introduced. This variable is YEAR. The difference between years can have quite severe consequences on the results of this research.

Introducing the year variable will help control the effect e.g. COVID-19 disaster year 2020 will have on the results of the entire dataset, events also known as "black swans". The incorporation of the year variable as a control variable will be as a dummy variable, given the fact that it is treated as a categorical variable.

3.3.3 Dependent variable

Table 5: Dependent variables and the methods of analysis

Variable	Abbreviation	Explanation
3-day cumulative abnormal return	CAR_3	The CAR_3 is concerned with measuring the cumulative abnormal return 3 trading days after issue of the annual report.
5-day cumulative abnormal return	CAR_5	The CAR_5 is concerned with measuring the cumulative abnormal return 5 trading days after issue of the annual report.
10-day cumulative abnormal return	CAR_10	The CAR_10 is concerned with measuring the cumulative abnormal return 10 trading days after issue of the annual report.

The CARs can be calculated by observing the actual stock return of a stock and subtracting the expected stock return for the same period. The reason for choosing the abnormal returns instead of the actual stock returns is to account for systematic risk as described in the control variables section. These abnormal returns will be cumulated for different time periods. For this research, these periods will be 3 days, 5 days, and 10 days. The expected stock return will be calculated according to the asset pricing model as identified in the literature, the CAPM (Pagliarussi et al., 2016; Duan et al., 2018):

$$E(r_{it}) = \beta_i * r_{mt} \tag{7}$$

Where $\mathrm{E(r_{it})}$ is the expected return of stock i in period t, β is a measure of volatility of stock i in relation to the market, and $\mathrm{r_{mt}}$ is the return of the market in period t. This formula, contrary to the original CAPM, does not include the risk-free rate. This can be explained by the fact that stock fluctuations are of importance for this research, not the annual required returns by investors.

As explained earlier in the section, the next step is to subtract the expected stock return from the actual stock return. In terms of calculation, this process would look like this (Pagliarussi et al., 2016; Duan et al., 2018; Yekini et al., 2016):

$$AR_{it} = r_{it} - E(r_{it}) \tag{8}$$

Where $\mathrm{AR_{it}}$ is the abnormal return of stock i in period t and $\mathrm{r_{it}}$ is the actual return of stock i in period t.

This step is followed by the accumulation of the abnormal returns generating the CAR, the CAR being the dependent variable to be used in the analysis:

$$CAR_{it} = \sum_{n=1}^{t} AR_{in} \tag{9}$$

Where CAR_{it} is the cumulative abnormal return of stock i over period t.
The retrieval and calculation of the different returns can be found in Appendix B.4 and B.5, respectively.

3.4 Data cleaning

Now that all data is processed, it needs to be "cleaned". In this research, this process consists of 2 steps: winsorization and normalization.

3.4.1 Winsorization

Within regression analysis, topics like measurement errors and outliers are a common issue. As established earlier in this section, this is the case for this research, as well. In order to tackle this issue, data can be cleaned in several ways, among which winsorization is a popular method (Lien & Balakrishnan, 2005). Winsorization has shown to increase the accuracy of predicted rates of returns of stock from historical rates of returns of the same stock. This can be explained by the fact that winsorization limits the extreme values in data, reducing the effect of outliers, which in terms of stock returns are often not very good predictors of future performance (Welch, 2017). Because of this observed effect of winsorization on especially stock return outliers, this technique will also be applied to the returns involved in this research, in order to diminish the effect of outliers. However, the role of winsorization does not stop there, since its application of mitigating the outlier sensitivity of textual data has also made winsorization a relevant tool in the field of sentiment analysis (Li, 2006; Tetlock, 2007). Hence, the technique of winsorization is also utilized for the independent variables of this thesis.

3.4.2 Normalization

Whenever there is a considerable difference between the ranges of variables within a research, the process of normalization is required (Ali, Faraj, & Koya, 2014). Since this is the case with this research, with variables like positivity, readability, and text density having considerably different ranges, normalization is applied to this research. It is important to note, however, that normalization is susceptible to outliers, making the previous process of winsorization the more important for the reliability of this research. For this research, a min-max normalization has been applied, meaning all variables have been normalized on a 0 to 1 scale. On top of the improved reliability for the regression analysis, normalization has also shown to improve the performance of machine learning models (Singh & Singh, 2020).

3.5 Regression analysis

The aim of this research, as described in the research question, is to research the effect of different types of annual report sentiment on stock returns. However, before this effect can be investigated, it is important to establish the significance of the cumulative abnormal returns that were observed. This excludes the possibility that these observations were gained by accident or as a result of a bias. This can be established via applying a t-test to the variables involved (Suwanna, 2012). After this step is completed, the actual regression analysis can start. For this research, a classic OLS (Ordinary Least Squares) regression is applied, as is the standard when researching stock returns (Joseph & Vezos, 2006). To investigate the effect the independent variables have on the dependent variable, the following regression model is fit to the data:

$$CAR\_X_{it=}\beta_0 + \beta_1 POS_i + \beta_2 REA_i + \beta_3 LIT_i + \beta_4 UNC_i + \beta_5 CON_i + \beta_6 SUP_i + \\ \beta_7 TEX_i + YEAR + e_{it} \tag{10}$$

Where $\mathrm{CAR\_X_{it}}$ are the cumulative abnormal returns of stock i for period t, β_0 is the intercept, $\beta_1\mathrm{POS_i}$ is the positivity variable for stock i, $\beta_2\mathrm{REA_i}$ is the readability variable for stock i, $\beta_3\mathrm{LIT_i}$ is the litigious variable for stock i, $\beta_4\mathrm{UNC_i}$ is the uncertainty variable for stock i, $\beta_5\mathrm{CON_i}$ is the constraining variable for stock i, $\beta_6\mathrm{SUP_i}$ is the superfluous variable for stock i, $\beta_7\mathrm{TEX_i}$ is the text density variable for stock i, YEAR is the year control variable, and $\mathrm{e_{it}}$ is the error term.

The variables POS, REA, LIT, UNC, CON, SUP, and TEX were, both with and without the control variable YEAR, regressed against the cumulative abnormal returns for 3-, 5-, and 10-day periods as described earlier. This resulted in a regression with 7 dependent variables and a control variable. However, regressions become increasingly complex with each added variable, which is why an additional model selection procedure was conducted to find the least complex but best fitting model. To find this model, Akaike's (1973) information criterion (AIC) is applied. The AIC gives a score to each model, indicating the "goodness of fit" of different models, allowing us to empirically find the best regression model. The technique is used in a plethora of fields, among which the analysis of stock returns (Domian & Louton, 1997; Miah, Rahman, et al., 2016). The AIC takes into account the maximum likelihood estimate of the model and the number of independent variables involved, thus its complexity. The lower the score, the better the fit of the model.

On top of that, for specific testing of the hypotheses formulated, all variables will individually, again with the control variable YEAR, be regressed against the 3-day cumulative abnormal returns, as is common when testing hypotheses (Li, 2010; Yekini et al., 2016; Hsieh, Hui, & Zhang, 2016). For the hypothesis testing only the 3-day cumulative abnormal return is used, because the contribution of this research is largest when relationships are established according to the most commonly used time period. When studying short-term abnormal returns, the commonly used time period is three days (Duan et al., 2018).

Complementary to the regression analysis, a Pearson's correlation matrix was made. This matrix shows the correlation between variables in degrees of strength of the correlation and the significance. Strong correlations are observed when the value is higher than 0.5 or lower than -0.5. Since this was the case during the research, the Variance

Inflation Factor (VIF) values for all analyzed models were measured as well. The VIF is a determinant of multicollinearity, a phenomena that can interfere with the significance of a model. When these VIF values exceed the cutoff point of 10, multicollinearity is a serious threat to the model (Craney & Surles, 2002). Multicollinearity in a model creates misleading results via the high correlations between independent variables, interfering with the effect of the individual independent variables on the model.

3.6 Supervised machine learning algorithm

On top of the regular research into the effects of sentiment on cumulative abnormal returns, a classification algorithm is created. For this classification algorithm several techniques are available. For this research, in line with relevant pieces of the prior literature in the field of classification algorithms with relation to stock returns (Tan, Yan, & Zhu, 2019; Lohrmann & Luukka, 2019), the random forest technique is utilized. The random forest model is designed as a "forest" of trees, with each tree predicting the classification of the dependent variable on the basis of the characteristics on the independent variables (Breiman, 2001). For these decision trees, there is no singular truth in the optimal number of trees utilized (Oshiro, Perez, & Baranauskas, 2012). This number is influenced by the goal and application of the algorithm, which will be elaborated on later in this section. In order to train these trees in their predicting capabilities, the dataset is split in a training and a testing set according to a 80:20 distribution. This is to say, 80% of the total dataset is utilized to train the machine learning algorithm in its classification capabilities. The remaining 20% of the dataset is utilized to test the classification capabilities learned by the algorithm on data that is "out-of-sample", so not part of the data utilized to train the algorithm. For this classification algorithm, the CAR_3, CAR_5, and CAR_10 are divided into positive and negative cumulative abnormal returns, with the cutoff point being at 0.

In order to quantify the testing of the algorithm 3 figures are of importance: (1) the accuracy, (2) the recall, and (3) the precision. The accuracy of the model is the percentage of observations classified correctly. This figure needs to be higher than the most obvious guess,or hypothesis probability, which would be the most occurring classification. The recall is the percentage of positives that are actually classified correctly as positives, and the precision is the percentage of observations classified as positive that are actually positive. For this research and its practical applicability, the recall is of the most importance. This can be explained when looking at the most important aspect of the algorithm: to identify stocks that are most likely to generate an abnormal positive return in the short-term. This goal makes it more important to classify positive return stocks correctly as positive return stocks than to have the positively classified stocks actually being positive return stocks. This is especially true given the fact that we assume investors are rational thinkers, making them a second line of defense in filtering out negative return stocks wrongly classified as positive, assuming the algorithm is not directly linked to an algorithmic trading bot.

Since it has been established that, from a practical standpoint, the most important statistic for the classification algorithm is the recall, the number of decision trees should be defined accordingly. Taking into account the fact that there is no general optimal number of trees, the optimal number of trees is research dependent, and that a large number of trees increases calculation times, an empirical method of establishing the optimal number of trees was used. This empirical method consisted of a mathematical optimization of the algorithm, making a "for" loop decide the most optimal number of trees. This optimization method mathematically decides what the most optimal number of trees is, resulting in the highest recall score for the 3 time period algorithms collectively, for which the code can be found in Appendix B.6.

Within classification algorithms there are correct predictions, type 1 errors, and type 2 errors. In line with the recall and precision of the algorithm are the 2 types of errors. The type 1 error is also known as a false positive, where an object is classified as positive without actually being positive, making it the opposite of precision. The type 2 error is also known as a false negative, where an object is classified as negative without actually being negative, making it the opposite of recall.

4 Results

4.1 Descriptive statistics

In table 6 the descriptive statistics for all relevant variables used in the regression models are shown, with the exception of the YEAR variable, which is treated as a categorical variable. The variables Pages and Words are included since they are the basis for the TEX variable. Most of the variables in these descriptive statistics need no further elaboration on the speciality of their values. However, certain variables in the table will be discussed in more detail in order to address important observations in their values.

Figures that call for extra attention in these descriptive statistics are the CARs, which mean and median values are close to zero. This is in line with Yekini et al. (2016), who state that the positive and negative reactions in stock markets cancel each other out, resulting in an expected average value of zero. Not only are these CAR descriptives in line with the statement made by Yekini et al. (2016), they also bear close resemblance to CARs in comparable studies (Rahman, 2019).

On top of the CARs, a noteworthy statistic is the negative median of POS. This statistic, possibly influenced by the skewed sizes of the POS dictionaries, indicates that annual reports are more likely to have a negative tone. This is a result that derives its noteworthiness from its contradicting value to what might be expected from such a clearly indicative variable.

Another variable that is of interest is REA, since its scale makes it an excellent variable to understand in terms of the implications of its values. That is to say, the range of annual reports analyzed vary in needed reading level from high school junior (12) to college graduate (17), whereas the average annual report required the reading level of a college sophomore (14). Since widely understood texts are often recommended to have a score of less than 12, it is clear to see the analyzed annual reports are not written in a way that makes them accessible to the wider public.

Lastly, it is of added value to discuss the implications of the scores of the variables LIT, UNC, CON, and SUP. These values can also be described as the share of the number of characteristic-related words in comparison to the total amount of words in processed annual reports. In practice, this shows that on average, after processing, 1 in every 55 words is related to uncertainty, whereas only 1 in every 114 words is related to constraining language. These statistics are for the texts after processing.

Table 6: Descriptive statistics

Variables	N	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
Pages	919	197.4	71.54	92	138	192	244	350
Words	919	90845	39609.48	39978	61228	82536	109352	191210
POS	919	-0.00521	0.00792	-0.0218	-0.105	-0.00391	0.000685	0.00622
REA	919	14.42	1.34	12.10	13.52	14.35	15.28	17.20
LIT	919	0.0110	0.00487	0.00522	0.00746	0.00981	0.0132	0.0237
UNC	919	0.0183	0.00453	0.0126	0.01467	0.01684	0.02126	0.0279
CON	919	0.00878	0.00222	0.00587	0.00713	0.00815	0.0100	0.0139
SUP	919	0.000434	0.000268	0.0000773	0.000203	0.000395	0.000618	0.00100
TEX	919	457.6	95.384	285.9	392.1	460	517.6	646.8
CAR_3	919	0.00231	0.0305	-0.0553	-0.0172	0.000696	0.0195	0.0674
CAR_5	919	0.00427	0.0413	-0.0726	-0.0220	0.00168	0.0280	0.0917
CAR_10	919	0.00825	0.0604	-0.0939	-0.0314	0.00157	0.0343	0.158

Note. Summary statistics after winsorization, before normalization. This choice was made to clearly illustrate the scales that were initially used in the analysis, e.g. the Gunning-Fog Index for the REA variable. All variables are defined in Table 3.

4.2 Variable measurement significance

As indicated when describing the regression analysis that is performed in this thesis, the checking of the significance of observations is of importance. In line with Suwanna (2012), the presence of cumulative abnormal returns is statistically checked in order to ensure the abnormality of the returns. For this, t-tests were performed on all relevant variables in the dataset, whose results can be seen in table 7. Checking all variables, and not only the CARs, further limits accidentality of observations in the entire width of the research. In this table 7, it is clear to see that all variables test significant for their t-statistic, assuring us that the observations are indeed not gained by accident. Aside from the fact that these t-tests enhance the reliability of this research, it shows that there are indeed cumulative abnormal returns to be observed after the issuance of an annual report by a listed company. This result is of critical value to the significance of this research.

Table 7: T-statistic of variables

	POS	REA	LIT	UNC	CON	SUP	TEX	CAR_3	CAR_5	CAR_10
T	63.56	52.75	36.24	38.11	39.664	40.28	54.57	57.37	56.42	51.26
df	918	918	918	918	918	918	918	918	918	918
p	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001

4.3 Assumptions

In order to conduct a successful OLS regression, four assumptions have to be fulfilled, these are; linearity, normality, homogeneity, and independence (Choi, 2002). The results for these assumptions can be found in figure 11 and their meaning will be elaborated on in the next sections.

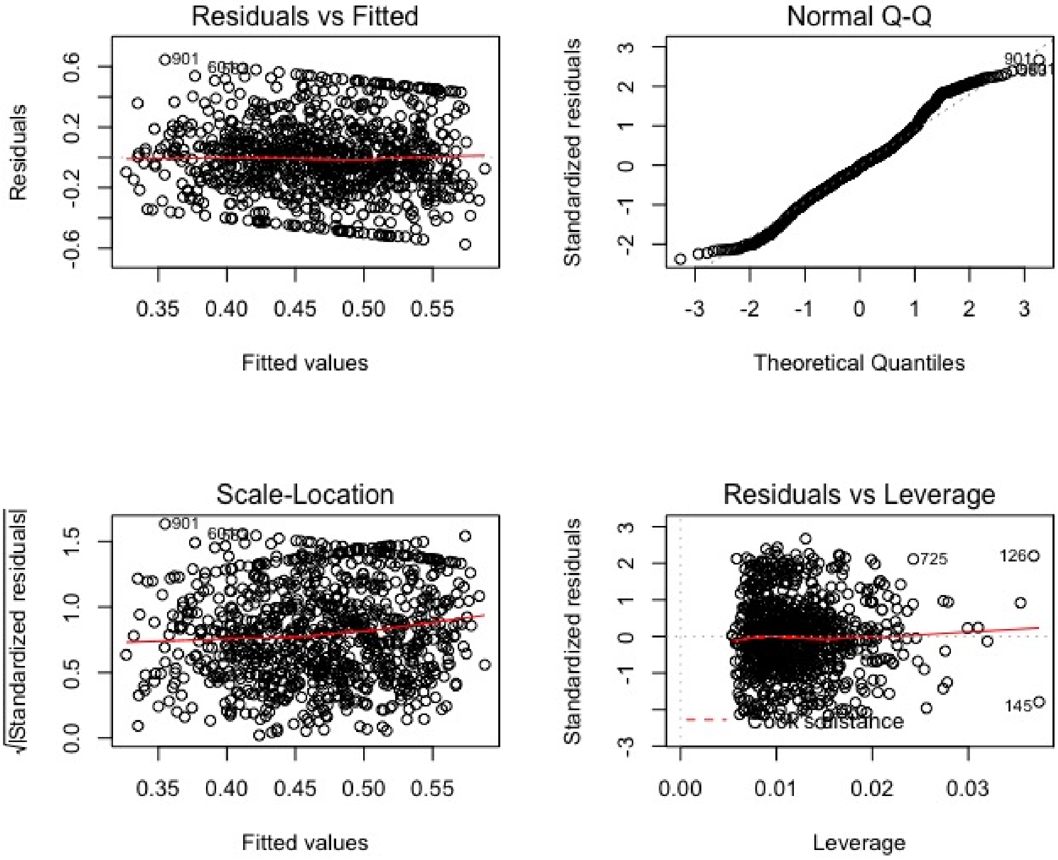


Figure 11: Assumptions

4.3.1 Linearity assumption

The linearity assumption is concerned with the fact that the relationship between the independent and the dependent variable is linear. This can be tested by plotting a Residuals vs Fitted graph which requires a horizontal line with no patterns in the spread of residuals for the assumption to be fulfilled. For this research, the linearity assumption is fulfilled since the Residuals vs Fitted consists of a clear horizontal line with the absence of patterns in the residuals.

4.3.2 Normality assumption

The normality assumption is, like the name suggests, concerned with the normal distribution of the data. This can be assumed on the basis of a large enough sample size (central limit theorem) or by plotting a Normal Q-Q figure. Within this figure the points plotted need to follow the dashed line running through the figure. The normality assumption is accepted via the central limit theorem, which states that sample sizes of 30 or larger tend to be sufficient to assume normality. On top of that, the Normal Q-Q figure confirms this, showing that the residual points follow the dashed line.

4.3.3 Homogeneity assumption

The homogeneity assumption, sometimes referred to as homoscedasticity, is concerned with the constant variance of residuals. This assumption can be tested by plotting the Scale-Location figure. In order to accept the homogeneity assumption, the Scale-Location needs to show a horizontal line with an equal spread in the points surrounding it. In this research the line, while bending upward slightly after 0.50, is sufficiently horizontal to assume homogeneity.

4.3.4 Independence assumption

This assumption is concerned with the independency of observations in the sample. While this from a logical viewpoint will be the case in this research, since observations are not influenced in any way by other observations as a characteristic of the research. On top of that, influential outliers were investigated with the Residuals vs Leverage plot, which does not show significant individual observations influencing the regression model from an outlier position.

4.4 Correlation

Within the independent variables in the dataset, a considerable number of significant correlations can be found. These correlations can be found in table 8. Of these significant correlations, some are of such a nature that further discussion of them is of interest to the outcomes of this research. In general, variables are said to be strongly correlated if the correlation is below -0.5 or above 0.5. Within this research, these values are observed between independent variables.

The variables that have strong correlations are POS, UNC, LIT, and CON. Of these correlations, it shows that the variable POS correlates in a strong but negative way with UNC, LIT and CON. Consequently, UNC, LIT, and CON correlate strongly positive with each other, which suggests that the effects of the POS variable (+) will be the opposite of the UNC, LIT, and CON variables (-). This is in line with what is to be expected from the literature for all variables except CON.

An important aspect to take into account with variables with high collinearity is multicollinearity. This phenomenon of multicollinearity is known to negatively influence the statistical significance of highly correlated independent variables (Allen, 1997). As a result of the high collinearity observed and discussed earlier in this section, it is important to consider this problem when conducting the regression analysis, taking this into account when choosing the independent variables to regress. In order to tackle this aspect of multicollinearity, the Variance Inflation Factor (VIF) for each model is added in Appendix C. These tables show no worrying VIF values, since the observed values in the models do not exceed 10, which is the cutoff value for multicollinearity.

Table 8: Pearson correlations between variables

	POS	REA	UNC	SUP	LIT	CON	TEX	CAR_3	CAR_5	CAR_10
POS	1.00									
REA	-0.23***	1.00								
UNC	-0.69***	0.12***	1.00							
SUP	0.14***	-0.36***	-0.11**	1.00						
LIT	-0.84***	0.23***	0.56***	-0.06	1.00					
CON	-0.77***	0.27***	0.60***	-0.21***	0.77***	1.00				
TEX	-0.17***	0.29***	0.22***	-0.27***	0.16***	0.22***	1.00			
CAR_3	0.13***	-0.07	-0.07	0.06	-0.10**	-0.08	-0.08	1.00		
CAR_5	0.11**	-0.08	-0.05	0.07	-0.06	-0.03	-0.06	0.80***	1.00	
CAR_10	0.07	0.00	-0.05	0.00	-0.03	-0.01	-0.02	0.58***	0.73***	1.00

Note. All variables are defined in Table 3. Significance of the Pearson's correlation test at the 95% and 99% confidence level is denoted as ** and ***, respectively.

4.5 Regression results

The regression results as described in table 9 contain the regression analysis when involving all independent variables as explanatory variables in the regression. In essence, these regressions show, all variables considered, what the explaining effect of the independent variables is on the cumulative abnormal returns. Within the regressions for the different time periods of 3-, 5-, and 10- days after the issuance of an annual report, several findings were reported.

First of all, the first result that catches the eye is the significance of the POS variable, the significance of which, in varying degrees of strength, is present in all time periods whether the control variable is omitted or included. This effect is positive among all regressions. The CON variable has a significant positive effect on CARs for 5- and 10-day time periods, while remaining insignificant at the 3-day level. This holds true for both the model with the control variable as well as for that without the control variable. Within the regressions there is one more significant factor, which is SUP in the controlled model for the 5-day cumulative abnormal returns. In that case, SUP returns a positive significant effect on the dependent variable. The 6 models presented in the table report varying results in terms of coefficients of determination, or R squared. In the table, it is clear to see the models which include the control variable of YEAR perform considerably better, as was expected when the choice for this control variable was made. The models including the control variable have explained variances ranging from 4% to 7.6%. These figures become larger as the time period is increased, as far as the controlled model is concerned.

As elaborated on in the methodology, regressions including all independent variables are often more complex than practically desirable when the goal is to create the best fitting regression model. Therefore, AIC was introduced in this research to evaluate the best fitting models on several criteria, giving each possible model a score. In figure 12, for illustrative purposes, the AIC scores of the 100 best fitting regression models for 3-day cumulative abnormal returns are shown.

Table 9: Results of regression analysis of sentiment analysis on 3-, 5-, and 10-day cumulative abnormal returns.

	With control variable			Without control variable		
	CAR_3	CAR_5	CAR_10	CAR_3	CAR_5	CAR_10
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	0.338***	0.238***	0.210***	0.361***	0.285***	0.286***
	(0.066)	(0.066)	(0.062)	(0.064)	(0.065)	(0.063)
POS	0.151**	0.203***	0.132**	0.169***	0.223***	0.148**
	(0.062)	(0.062)	(0.059)	(0.062)	(0.063)	(0.061)
REA	-0.025	-0.040	0.011	-0.026	-0.043	0.005
	(0.035)	(0.035)	(0.033)	(0.035)	(0.035)	(0.034)
LIT	-0.038	0.006	0.003	-0.007	0.047	0.047
	(0.061)	(0.061)	(0.058)	(0.061)	(0.062)	(0.060)
UNC	0.033	0.024	-0.008	0.031	0.022	-0.010
	(0.039)	(0.039)	(0.037)	(0.039)	(0.040)	(0.038)
CON	0.081	0.140***	0.101**	0.073	0.127**	0.084*
	(0.051)	(0.051)	(0.048)	(0.051)	(0.052)	(0.050)
SUP	0.034	0.055*	0.011	0.024	0.041	-0.003
	(0.031)	(0.031)	(0.029)	(0.031)	(0.032)	(0.030)
TEX	-0.038	-0.014	-0.007	-0.054	-0.032	-0.021
	(0.033)	(0.033)	(0.032)	(0.033)	(0.034)	(0.032)
N	919	919	919	919	919	919
Control	Included	Included	Included	Omitted	Omitted	Omitted
Residual SE	0.243	0.243	0.231	0.246	0.249	0.240
F-statistic	4.788	7.144	8.522	3.416	3.644	1.385
Adjusted R^2	0.040	0.063	0.076	0.018	0.020	0.003

Note. Regression analysis where the cumulative abnormal returns for 3-, 5-, and 10-day time periods act as dependent variables. All variables are defined in Table 3. Standard errors of the coefficient estimates are given in parentheses. Significance of the coefficient estimates at the 90%, 95%, and 99% confidence level is denoted as *, **, and ***, respectively.

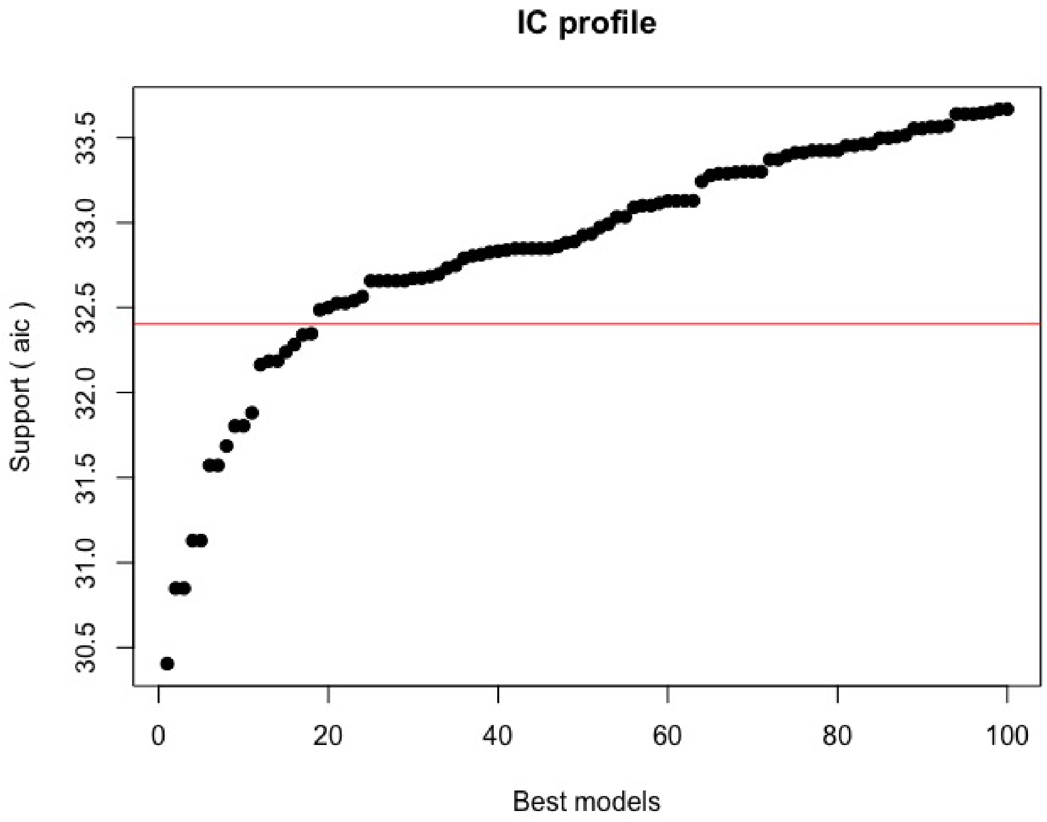


Figure 12: AIC of models for 3-day CARs

As a result of these AIC scores, new regression analyses were performed. These new analyses resulted in less complex but comparable or better performing regression models, which can be found in table 10. On top of the models only regarding main effects, interaction effects between the variables were studied. As can be seen from the table, these models oftentimes included the CUL variable as an extra variable, showing the difference the culture of the country of listing can make to the regression model. The addition of interaction effects, which are often involved with the CUL variable, has shown to improve the 3-, 5-, and 10-day CAR models by 0.4%, 0.6%, and 1.4% respectively.

Table 10: Results of regression analysis of sentiment analysis on 3-, 5-, and 10-day cumulative abnormal returns, based on AIC scores.

	CAR_3		CAR_5		CAR_10	
	(1)	(2)	(1)	(2)	(1)	(2)
Intercept	0.307***	0.327***	0.227***	0.230***	0.215***	0.192***
	(0.047)	(0.038)	(0.047)	(0.047)	(0.042)	(0.049)
POS	0.156***	0.134***	0.189***		0.134***	
	(0.045)	(0.044)	(0.045)		(0.042)	
REA						
UNC						
CON	0.067		0.139***	0.134***	0.099**	0.085*
	(0.046)		(0.046)	(0.046)	(0.043)	(0.044)
SUP	0.047*		0.070**			
	(0.028)		(0.028)			
CON·UNC		0.087*				
		(0.051)				
POS·CUL		0.062**				
		(0.025)				
CUL·POS				0.143***		0.115**
				(0.047)		(0.047)
CUL·UNC						0.071*
						(0.041)
CUL·SUP				0.127***		
				(0.046)		
N	919	919	919	919	919	919
Control	Included	Included	Included	Included	Included	Included
Residual SE	0.243	0.242	0.243	0.242	0.230	0.229
F-statistic	7.401	8.091	11.56	9.739	17.06	12.88
Adjusted R^2	0.040	0.044	0.065	0.071	0.080	0.094

Note. Regression analysis where the cumulative abnormal returns for 3-, 5-, and 10-day time periods act as dependent variables. The models (1) are without interaction effects. The models (2) are with interaction effects. All variables are defined in Table 3. Standard errors of the coefficient estimates are given in parentheses. Significance of the coefficient estimates at the 90%, 95%m, and 99% confidence level is denoted as *, **, and ***, respectively.

4.6 Hypothesis testing

In order to test the hypotheses that were constructed after consultation of the literature, each individual independent variable has been regressed against the 3-day cumulative abnormal returns. As can be seen from Table 10, all independent variables that were hypothesized upon have significant effects on the 3-day cumulative abnormal returns. This means all effects as portrayed in the table and the remainder of this section are significant in varying degrees, so this will not be mentioned again at every individual variable. The observed significant effects, however, were in some cases not what was expected when the hypotheses were formulated. Therefore, the variables will be individually discussed.

First of all, the most influential variable of POS reports a positive effect on CARs. This is in line with the formulated hypotheses and the literature as derived from sources like Wisniewski & Yekini (2015) and Azimi & Agrawal (2021).

The second variable which was researched was REA, which denoted a clear negative effect on CARs. This result was in line with the formulated hypotheses and the researched literature, among which Li (2018) and Miller (2010).

Thirdly, the effect of the variable LIT was studied. This effect turned out to be negative, which was, once again, in line with the formulated hypotheses and the underlying literature as derived from Pagliarussi et al. (2016) and Lim et al. (2018).

The fourth variable of interest was UNC. This variable, however less significant than most of the other variables, showed a negative effect on CARs. This result supports the formulated hypotheses and indirect sources from the literature utilized to formulate this hypothesis, Ozoguz (2009).

Fifth of all, the variable CON and its effect on 3-day CARs were researched. In contrast to the formulated hypothesis, which predicted a positive effect, a negative effect on CARs was found. Consequently, this finding is also not in line with the literature utilized as benchmark for the formulated hypothesis, Buehlmaier & Whited (2015).

The sixth variable SUP, and its corresponding hypothesis of a negative effect of SUP on 3-day CARs, were found to have a significant positive effect on CARs. This result is not in line with the expected result formulated in the hypothesis, which derived from the relationship between legislation and superfluous language in annual reports.

Lastly, the effect of TEX on CARs were studied. This effect, as was proposed in the corresponding hypothesis, is negative and, therefore, supports the expected results of this effect.

Table 11: Effect of independent variables on 3-day cumulative abnormal returns.

Variables	Hypothesis	Sign	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept			0.376*** (0.0241)	0.477*** (0.022)	0.476*** (0.019	0.464*** (0.019)	0.468*** (0.019)	0.424*** (0.020)	0.474*** (0.021)	0.338*** (0.066)
POS	H1	+	0.111*** (0.029)							0.151** (0.062)
REA	H2	-		-0.069** (0.031)						-0.025 (0.035)
LIT	H3	-			-0.101*** (0.031)					-0.038 (0.061)
UNC	H4	-				-0.053* (0.027)				0.033 (0.039)
CON	H5	+					-0.065** (0.029)			0.081 (0.051)
SUP	H6	-						0.054* (0.028)		0.034 (0.033)
TEX	H7	-							-0.062** (0.031)	-0.038 (0.033)
N			919	919	919	919	919	919	919	919
Residual SE			0.243	0.245	0.244	0.245	0.245	0.245	0.245	0.243
F-statistic			10.04	7.400	8.900	7.088	7.400	7.087	7.159	7.788
Adjusted R^2			0.038	0.027	0.033	0.026	0.027	0.026	0.026	0.040

Note. Hypothesis testing via linear regression analysis, where the cumulative abnormal returns for the 3-day time period act as dependent variable. All variables are defined in Table 3. The hypothesis sign of REA is flipped, since a higher score means lower readability. Standard errors of the coefficient estimates are given in parentheses. Significance of the coefficient estimates at the 90%, 95%, and 99% confidence level is denoted as *, **, and ***, respectively.

4.7 Supervised machine learning algorithm

As described in the methodology, a random forest machine learning algorithm was trained on the dataset. This machine learning algorithm has been trained to classify annual report sentiment inputs as positive or negative expected cumulative abnormal stock returns. For this machine learning algorithm 197 decision trees were used, since this was the number that was decided as optimal for the applications of this algorithm specifically, by means of mathematical optimization. The relevant scores of the algorithm can be found in table 12 and were retrieved by utilizing the code found in Appendix B.7. In short, it is clear to see the machine learning algorithm performs better than the highest probable outcome from the data, the hypothesis probability, which would in this case be a ”positive return”. On average, the algorithm is more accurate than the highest probable outcome from the data by 6.1%. Also, the most important statistic of this algorithm, the recall, reports scores of 0.674, 0.650, and 0.612 for the CAR_3, CAR_5, and CAR_10, respectively. This means that for every 1000 out-of-sample annual reports with a positive abnormal return on the 3 days after issue, 674 are correctly classified. In order to visualize the results of a classification algorithm, confusion matrices are created which show: the correct predictions, the type 1, and type 2 errors. For the CAR_3, this confusion matrix can be found in figure 13. For the other CAR predictions, the confusion matrices are included in appendix D.

Table 12: Performance of the random forest model

	Hypothesis probability	Accuracy	Recall	Precision
CAR_3	0.509	0.600	0.674	0.598
CAR_5	0.514	0.560	0.650	0.586
CAR_10	0.509	0.554	0.612	0.577

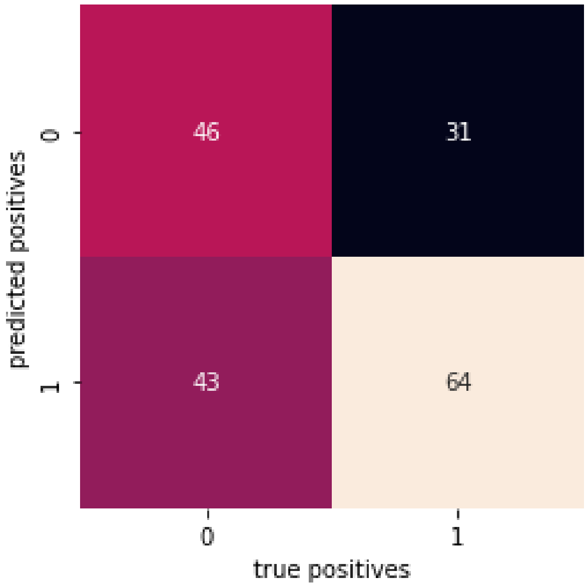


Figure 13: Confusion matrix for CAR_3 classification algorithm

5 Discussion and Conclusion

5.1 Key findings

The aim of this research is to find an answer to the research question: "How does the sentiment of an annual report influence short-term cumulative abnormal stock returns in the US, UK, Germany and the Netherlands?". For this, types of sentiment of approximately 900 annual reports from these countries were analyzed and utilized as explanatory variables for the cumulative abnormal returns for these companies the days after annual report issuance. Based on the findings of this study, several conclusions can be drawn.

First of all, it can be concluded that, when taking them into account individually, all types of sentiment researched have an effect on cumulative abnormal returns. These effects are in most cases in line with prior literature, like is the case for the positive effect of positive annual reports on CARs, the negative effect of low readability in annual reports on CARs and the negative effect of a high amount of litigious text in an annual report on CARs. In other cases, the effects of the independent variables were in line with what was expected, but these expectations have derived from indirect literature, in light of the absence of direct underlying literature. Therefore, the negative effect of uncertainty in annual reports and the negative effect of higher text densities on CARs are new additions to the literature, deriving from the individual analysis of these variables against short-term cumulative abnormal returns.

Not all variables showed the effects that were preliminary expected, though. The negative effects of constraining language and the positive effects of superfluous language in annual reports on short-term CARs were counter-intuitive conclusions from this research. However, these counter-intuitive conclusions are nevertheless an important addition to the existing literature, shedding a new light on the effects of constraining and superfluous language.

When the individual variables are considered together as an influence on short-term cumulative abnormal returns, it showed that some variables had stronger influences than others. It is important to mention the strong and continuous positive effect positivity remained to have on the cumulative abnormal returns in any of the chosen time periods. On top of the positivity effect, the positive effect constraining language has on cumulative abnormal returns when considering any but the shortest of time periods is an important finding for the literature, especially when taking into account its contradicting result in relation to the prior literature.

From a more practical standpoint, it can be concluded that the sentiment of annual reports is an adequate way of improving decision-making for investors. From the results it became clear that, when trained on annual report sentiment, a machine learning algorithm is capable of classifying stock prices of companies issuing annual reports into positive and negative abnormal return stocks. The algorithm is reported to be more accurate, have better recall, and have better precision than the hypothesis in all time periods.

Summarized, the results of this research contribute to the literature in 2 important ways. First, by showing the impact cultural considerations have on the perception of annual report sentiment, in line with developments in other fields such as corporate governance (Aguilera & Crespi-Cladera, 2016). Having done this, the research expands the field of research, showing the added value of transitioning from the current singular-country, more national approach, to international approaches, accounting for cultural differences between markets. Secondly, this research shows that the combination of the chosen variables are a better explanatory fit for the variance in cumulative abnormal stock returns. Not only does it trump the unilateral research efforts taking into account one type of sentiment, which reported explained variances below the 1% mark (Yekini et al., 2016), it also takes a significant leap forward from the explained variances of research efforts taking into account a combination of sentiment types, which often hover anywhere between 3 and 5% (Wisniewski & Yekini, 2015; Pagliarussi et al., 2016; Rahman, 2019). With explained variances exceeding the 9% mark for 10-day cumulative abnormal returns, this research should be seen as a significant step forward in the field of sentiment analysis as a predictor for short-term stock returns.

5.2 Limitations

Within this study, there are several limitations. The first limitation is concerned with the sample, which is limited in terms of the countries and years involved. Since the culture of the country in which the company is primarily listed has been proven to have quite the interaction effect, it is a limitation that the entirety of Continental European culture is based on Germany and the Netherlands. This small sample size in terms of countries does not account for effects in other countries with considerable stock exchanges, such as France. On top of that, the number of years considered, which is 4, is less than that of similar research, which oftentimes starts as low as at 6 years (Rahman, 2019). This limited number of years, even though accounted for via control variables, could fail to eliminate the effects rare macro-economic events have on the results.

Lastly, the practical application of the regression analysis is to be questioned. Empirically, it is correct that the regression is a step forward from earlier research efforts and the classification algorithm is a substantial improvement over the most educated guess. However, in order to accurately explain the variance in short-term abnormal stock returns from the investor's perspective, the explained variance not surpassing the 10% is still a limitation to this research and the literature in general. Even though it might help investors in practice, a 10% explained variance is not a measure to rely on on its own.

5.3 Future research

Deriving from this research, areas for further research have risen up that need to be addressed in a plethora of fields, of which 3 are catching the eye.

Firstly, within this research, and in the field in general, a lot of the more classic dictionary approaches are used. Nowadays, there are more advanced techniques available that, when finance and computer science get combined, can shed new lights on the topic of sentiment analysis in corporate filings, and the corresponding effects of it on stock markets. These techniques could help improve the accuracy of analyzed sentiment, taking into account not only individual words, but also the context for large sample sizes in terms of page and document count. These techniques are already used on a larger scale, like the research by Wujec (2021), however have not yet been applied on a serious scale to explanatory research involving the effect on corporate filings on abnormal stock returns.

Secondly, future research should widen the scope from national research to more globalized research. Topics like this could entail, but are not limited to, the cultural differences between Western countries and third-world countries; and the resulting effects of these cultural differences. This research has set the first steps into incorporating these variables into the equation from an explanatory perspective. On top of that, research should focus on solidifying the understanding we have of more closely linked cultures, like the Anglo-Saxon and Western European culture, creating more understanding of differences between cultures that oftentimes are considered quite similar.

Lastly, future research should focus on combining the sentiment of annual report filings and the corresponding news sentiment deriving from the annual report issuance. As became apparent during analysis of the results and the prior literature; both annual report and news sentiment have statistically significant but individually limited effects on short-term cumulative abnormal returns. Future research should focus on combining the sentiment of annual reports with the sentiment of news issues surrounding the annual report, trying to create a stronger understanding of not only the public perception of annual reports, but also the cause of significant abnormal returns surrounding annual report issues.

A Techniques for sentiment analysis

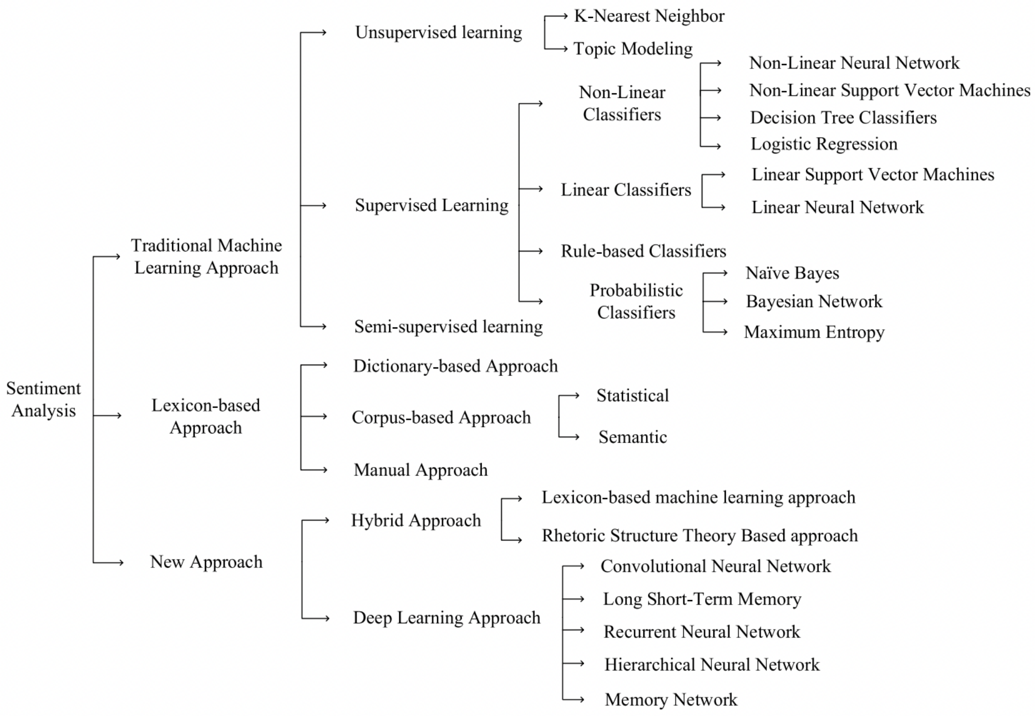


Figure 14: More extensive overview of sentiment analysis techniques (Shi et al., 2019)

B Code listings

B.1 PDF to TXT converter (AHK)

```
SourceFolder:="Filepath PDFs" ;
PDFtoTextEXE:="Filepath programme" ;
OutputFormat:="-raw" ;
Errors:=""
NumConverted:=0
If (SubStr(SourceFolder,0,1)!="\")
  SourceFolder:=SourceFolder . "\"
Loop,Files,%SourceFolder%*.pdf
{
  PDFfile:=A_LoopFilePath
  RunWait,"%PDFtoTextEXE%" %OutputFormat% "%PDFfile%",,Hide
  If (ErrorLevel!=0)
    Errors:=Errors . PDFfile . "‘n"
  Else
    NumConverted:=NumConverted+1
}
If (Errors="")
```

```
  Errors:="None"
MsgBox,4096,Done Converting,Number converted:
%NumConverted%`n`nErrors converting:`n%Errors%
ExitApp
```

B.2 Readability analysis (Python)

```
import pandas as pd

df = pd.read_excel('list of txt filenames')
mylist = df['Files'].tolist()

from textstat.textstat import textstat

for x in mylist:
    text = open('Filepath to txt files{}'.format(x),'r', encoding='latin-1').read()
    print(textstat.gunning_fog(text))
```

B.3 Dictionary analysis (Python)

```
import pandas as pd
import re
import nltk
from nltk.tokenize import RegexpTokenizer
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

df = pd.read_excel('List of txt filenames')
mylist = df['Files'].tolist()
for x in mylist:
    file = open('txt filepath/{}'.format(x), 'r', encoding = 'latin-1')
    text = file.read()
    text = text.replace('\n', ' ')
    text = text.replace('\x0c', ' ')
    text = re.sub('[^a-zA-Z]+', ' ', text).lower()
    negative = open('Negative dictionary', 'r', encoding = 'latin-1')
    negative = negative.read()
    positive = open('Positive dictionary', 'r', encoding = 'latin-1')
    positive = positive.read()
    litigious = open('Litigious dictionary', 'r', encoding = 'latin-1')
    litigious = litigious.read()
    uncertainty = open('Uncertainty dictionary', 'r', encoding = 'latin-1')
    uncertainty = uncertainty.read()
    constraining = open('Constraining dictionary', 'r', encoding = 'latin-1')
    constraining = constraining.read()
    superfluous = open('Superfluous dictionary', 'r', encoding = 'latin-1')
    superfluous = superfluous.read()
```

```
negative = negative.replace('\n', ' ')
negative = negative.lower().split()
positive = positive.replace('\n', ' ')
positive = positive.lower().split()
litigious = litigious.replace('\n', ' ')
litigious = litigious.lower().split()
uncertainty = uncertainty.replace('\n', ' ')
uncertainty = uncertainty.lower().split()
constraining = constraining.replace('\n', ' ')
constraining = constraining.lower().split()
superfluous = superfluous.replace('\n', ' ')
superfluous = superfluous.lower().split()
count_negative = 0
count_positive = 0
count_litigious = 0
count_uncertainty = 0
count_constraining = 0
count_superfluous = 0
for word in text.split():
    if word in negative:
        count_negative = count_negative + 1
    if word in positive:
        count_positive = count_positive + 1
    if word in litigious:
        count_litigious = count_litigious + 1
    if word in uncertainty:
        count_uncertainty = count_uncertainty + 1
    if word in constraining:
        count_constraining = count_constraining + 1
    if word in superfluous:
        count_superfluous = count_superfluous + 1
stop_words = set(stopwords.words('english'))
word_tokens = word_tokenize(text)
text1 = [w for w in word_tokens if not w.lower() in stop_words]
text1 = []
for w in word_tokens:
    if w not in stop_words:
        text1.append(w)
total_words = len(text1)
rel_negative = count_negative / total_words
rel_positive = count_positive / total_words
rel_litigious = count_litigious / total_words
rel_uncertainty = count_uncertainty / total_words
rel_constraining = count_constraining / total_words
rel_superfluous = count_superfluous / total_words
tone = rel_positive + rel_negative*(-1)
```

```
    print(tone,',',rel_litigious,',',rel_uncertainty,
    ',',rel_constraining,',',rel_superfluous)
print('done')
```

B.4 Retrieval of actual stock returns (R)

```
getSymbols("Country"$Ticker, from = '2017-01-01',
           to = "2021-12-31",warnings = FALSE,
           auto.assign = TRUE)

mylist <- unique("Country"$Ticker)

for (x in mylist) {
    "Country"$Return_3 <- get(x)$Change["Country"$Over] +
    get(x)$Change["Country"$Over+1] + get(x)$Change["Country"$Over+2]
```

B.5 Calculation of expected returns (R)

```
"Country"$Exp_3 <-  "Country"$Beta * ("Index"$Change["Country"$Over] +
"Index"$Change["Country"$Over+1] + "Index"$Change["Country"$Over+2])
```

B.6 Optimal number of decision trees for classifier (Python)

```
import seaborn as sns
import pandas as pd

Data = pd.read_csv(r'Filepath')
X = Data[["Positivity1", "Readability1", "Uncertainty1", etc.]]
X = X.values
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
tes={}
for x in range(1,300):
    tes[x]=0
    for i in ["CAR_3P","CAR_5P","CAR_10P"]:
        y = Data[[i]]
        y = y.values.ravel()
        X_train, X_test, y_train, y_test = train_test_split(X, y,
        test_size = 0.20, random_state = 116)
        np.random.seed(116)
        clf=RandomForestClassifier(n_estimators=x, random_state = 116)
        clf.fit(X_train,y_train)
        y_pred=clf.predict(X_test)
        tes[x]+=metrics.recall_score(y_test, y_pred, pos_label="Positive")
```

```
print(max(tes,key=tes.get))

Output: 197
```

B.7 Classification algorithm results (Python)

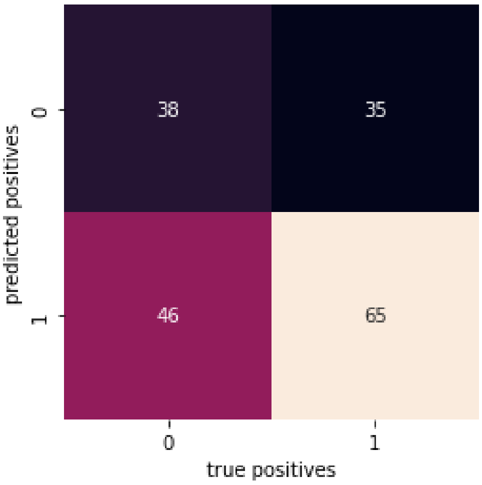
```
for i in ["CAR_3P","CAR_5P","CAR_10P"]:
    y = Data[[i]]
    y = y.values.ravel()
    X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size = 0.20, random_state = 116)
    np.random.seed(116)
    clf=RandomForestClassifier(n_estimators=197, random_state = 116)
    clf.fit(X_train,y_train)
    y_pred=clf.predict(X_test)
    print(metrics.recall_score(y_test, y_pred, pos_label="Positive"))
```

C VIF

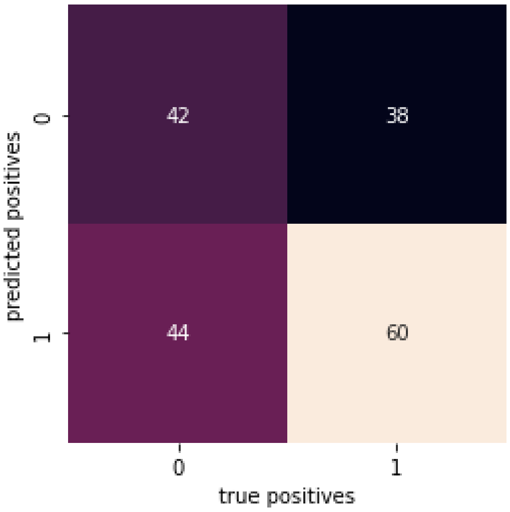
Table 13: VIF values regression model

	CAR_3		CAR_5		CAR_10	
	(1)	(2)	(1)	(2)	(1)	(2)
POS	2.480	2.428	2.480		2.477	
REA						
UNC						
CON	2.530		2.530	2.596	2.465	2.639
SUP	1.047		1.047			
CON*UNC		2.387				
POS*CUL		1.263				
CUL*POS				1.011		4.784
CUL*UNC						3.265
CUL*SUP				2.596		

D Confusion matrices



(a) Matrix for 5-day CARs



(b) Matrix for 10-day CARs

Figure 15: Matrices for classification algorithm

References

Aguilera, R. V., & Crespi-Cladera, R. (2016). Global corporate governance: On the relevance of firms' ownership structure. *Journal of World Business*, *51*(1), 50–57.

Ajina, A., Laouiti, M., & Msolli, B. (2016). Guiding through the fog: does annual report readability reveal earnings management? *Research in International Business and Finance*, *38*, 509–516.

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Ali, P. J. M., Faraj, R. H., & Koya, E. (2014). Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, *1*(1), 1–6.

Allen, M. P. (1997). The problem of multicollinearity. *Understanding regression analysis*, 176–180.

Aydogdu, M., Saraoglu, H., & Louton, D. (2019). Using long short-term memory neural networks to analyze sec 13d filings: A recipe for human and machine interaction. *Intelligent Systems in Accounting, Finance and Management*, *26*(4), 153–163.

Azimi, M., & Agrawal, A. (2021). Is positive sentiment in corporate annual reports informative? evidence from deep learning. *The Review of Asset Pricing Studies*, *11*(4), 762–805.

Bai, X., Dong, Y., & Hu, N. (2019). Financial report readability and stock return synchronicity. *Applied Economics*, *51*(4), 346–363.

Baker, M., & Wurgler, J. (2007a). Investor sentiment in the stock market. *Journal of economic perspectives*, *21*(2), 129–152.

Baker, M., & Wurgler, J. (2007b, June). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129-152. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/jep.21.2.129` doi: 10.1257/jep.21.2.129

Balata, P., & Breton, G. (2005). Narratives vs numbers in the annual report: are they giving the same message to the investors? *Review of Accounting and Finance*.

Biddle, G. C., Hilary, G., & Verdi, R. S. (2009). How does financial reporting quality relate to investment efficiency? *Journal of accounting and economics*, *48*(2-3), 112–131.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, *50*(4), 623–646.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative research journal*.

Brealey, R. A., Myers, S. C., Allen, F., & Mohanty, P. (2018). *Principles of corporate finance, 12/e* (Vol. 12). McGraw-Hill Education.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brownlee, J. (2017). Deep learning for natural language processing. *Machine Learning Mastery*, 414.

Buehlmaier, M., & Whited, T. M. (2015). Looking for risk in words: A narrative approach to measuring the pricing implications of financial constraints. In *Annual conference of the western finance association (wfa)*.

Cagnetti, A. (2002). Capital asset pricing model and arbitrage pricing theory in the italian stock market: an empirical study.

Chakraborty, M., & Subramaniam, S. (2020). Asymmetric relationship of investor sentiment with stock return and volatility: evidence from india. *Review of Behavioral Finance*.

Che, S., Zhu, W., & Li, X. (2020). Anticipating corporate financial performance from ceo letters utilizing sentiment analysis. *Mathematical Problems in Engineering*, *2020*.

Chin, W. W., et al. (1998). The partial least squares approach to structural equation modeling. *Modern methods for business research*, *295*(2), 295–336.

Choi, I. (2002). Econometrics: by fumio hayashi, princeton university press, 2000. *Econometric Theory*, *18*(4), 1000–1006.

Coşkun, Y., Selcuk-Kestel, A. S., & Yilmaz, B. (2017). Diversification benefit and return performance of reits using capm and fama-french: Evidence from turkey. *Borsa Istanbul Review*, *17*(4), 199–215.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality engineering*, *14*(3), 391–403.

Dhankar, R. S., & Singh, R. (2005). Arbitrage pricing theory and the capital asset pricing model-evidence from the indian stock market. *Journal of Financial Management & Analysis*, *18*(1), 14.

Domian, D. L., & Louton, D. A. (1997). A threshold autoregressive analysis of stock returns and real economic activity. *International Review of Economics & Finance*, *6*(2), 167–179.

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings conference call content and stock price: the case of reits. *The Journal of Real Estate Finance and Economics*, *45*(2), 402–434.

Duan, J., Zhang, Y., Ding, X., Chang, C. Y., & Liu, T. (2018). Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2823–2833).

Falk, R. F., & Miller, N. B. (1992). *A primer for soft modeling.* University of Akron Press.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*(4), 915–953.

Ferguson, N. J., Philip, D., Lam, H., & Guo, J. M. (2015). Media content and stock returns: The predictive power of press. *Multinational Finance Journal*, *19*(1), 1–31.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157–214.

Gandía, J. L., & Huguet, D. (2021). Textual analysis and sentiment analysis in accounting: Análisis textual y del sentimiento en contabilidad. *Revista de Contabilidad-Spanish Accounting Review*, *24*(2), 168–183.

Gangolly, J. S., Hedley, T. P., & Wong, C.-T. (1991). Semantic knowledge bases for financial accounting standards. *Expert Systems with Applications*, *3*(1), 117–128.

Hájek, P. (2018). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Computing and Applications*, *29*(7), 343–358.

Hájek, P., Olej, V., & Myskova, R. (2013). Forecasting stock prices using sentiment information in annual reports-a neural network and support vector regression approach. *WSEAS Transactions on Business and Economics*, *10*(4), 293–305.

Hamao, Y. (1988). An empirical examination of the arbitrage pricing theory: Using japanese data. *Japan and the World economy*, *1*(1), 45–61.

Hearst, M. (2003). What is text mining. *SIMS, UC Berkeley*, *5*.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, *45*(4), 363–407.

Henry, E., & Leone, A. J. (2010). Measuring qualitative information in capital markets research. *Papel de trabajo*.

Hsieh, C.-C., Hui, K. W., & Zhang, Y. (2016). Analyst report readability and stock returns. *Journal of Business Finance & Accounting*, *43*(1-2), 98–130.

Iquiapaza, R. A., Carneiro, R. L. d., Amaral, H. F., & Ferreira, B. P. (2021). Performance of active index stock funds using the capm from 1997 to 2019.

Jeon, Y., McCurdy, T. H., & Zhao, X. (2021). News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies. *Journal of Financial Economics*.

Jiang, S., & Jin, X. (2021). Effects of investor sentiment on stock return volatility: A spatio-temporal dynamic panel model. *Economic Modelling*, *97*, 298–306.

Joseph, N. L., & Vezos, P. (2006). The sensitivity of us banks' stock returns to interest rate and exchange rate changes. *Managerial Finance*.

Karppinen, K., & Moe, H. (2019). Texts as data i: Document analysis. In *The palgrave handbook of methods for media policy research* (pp. 249–262). Springer.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171–185.

Kendall, M. G., & Hill, A. B. (1953). The analysis of economic time-series-part i: Prices. *Journal of the Royal Statistical Society. Series A (General)*, *116*(1), 11–34.

Khin, E., Tee, L. K., & Ying, C. W. (2011). Cumulative abnormal returns on share buy back: Malaysian perspectives. *Australian Journal of Basic and Applied Sciences*, *5*(12), 2168–2175.

Kim, C., Wang, K., & Zhang, L. (2019). Readability of 10-k reports and stock price crash risk. *Contemporary accounting research*, *36*(2), 1184–1216.

Kim, K., Ryu, D., & Yang, H. (2019). Investor sentiment, stock returns, and analyst recommendation changes: The kospi stock market. *Investment Analysts Journal*, *48*(2), 89–101.

Kisman, Z., & Restiyanita, S. (2015). M. the validity of capital asset pricing model (capm) and arbitrage pricing theory (apt) in predicting the return of stocks in indonesia stock exchange. *American Journal of Economics, Finance and Management*, *1*(3), 184–189.

Lehavy, R., Li, F., & Merkley, K. (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review*, *86*(3), 1087–1115.

Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports? *Available at SSRN 898181*.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, *45*(2-3), 221–247.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, *48*(5), 1049–1102.

Lien, D., & Balakrishnan, N. (2005). On regression analysis with data cleaning via trimming, winsorization, and dichotomization. *Communications in Statistics—Simulation and Computation®*, *34*(4), 839–849.

Lim, E. K., Chalmers, K., & Hanlon, D. (2018). The influence of business strategy on annual report readability. *Journal of Accounting and Public Policy*, *37*(1), 65–81.

Lohrmann, C., & Luukka, P. (2019). Classification of intraday s&p500 returns with a random forest. *International Journal of Forecasting*, *35*(1), 390–407.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, *66*(1), 35–65.

Loughran, T., & McDonald, B. (2014). Regulation and financial disclosure: The impact of plain english. *Journal of Regulatory Economics*, *45*(1), 94–113.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.

Loughran, T., & McDonald, B. (2017). The use of edgar filings by investors. *Journal of Behavioral Finance*, *18*(2), 231–248.

Malkiel, B. G. (1989). Efficient market hypothesis. In *Finance* (pp. 127–134). Springer.

Martani, D., & Khairurizka, R. (2009). The effect of financial ratios, firm size, and cash flow from operating activities in the interim report to the stock return. *Chinese Business Review*, *8*(6), 44.

Miah, M., Rahman, A., et al. (2016). Modelling volatility of daily stock returns: Is garch (1, 1) enough? *American Academic Scientific Research Journal for Engineering, Technology, and Sciences*, *18*(1), 29–39.

Miller, B. P. (2010). The effects of reporting complexity on small and large investor trading. *The Accounting Review*, *85*(6), 2107–2143.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with python: a guide for data scientists.* " O'Reilly Media, Inc.".

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition* (pp. 154–168).

Ozoguz, A. (2009). Good times or bad times? investors' uncertainty and stock returns. *The Review of Financial Studies*, *22*(11), 4377–4422.

Pagliarussi, M. S., Aguiar, M. O., & Galdi, F. C. (2016). Sentiment analysis in annual reports from brazilian companies listed at the bm&fbovespa. *Base Revista de Administração e Contabilidade da UNISINOS*, *13*(1), 53–64.

Pajuste, A., Poriete, E., & Novickis, R. (2020). Management reporting complexity and earnings management: evidence from the baltic markets. *Baltic Journal of Management*.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Penrose, J. M. (2008). Annual report graphic use: a review of the literature. *The Journal of Business Communication (1973)*, *45*(2), 158–180.

Poshalian, S., & Crissy, W. J. (1952). Corporate annual reports are difficult, dull reading, human interest value low, survey shows. *Journal of Accountancy (pre-1986)*, *94*(000002), 215.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, *36*(4), 992–1011.

Qu, Z., Liu, X., & He, S. (2019). Abnormal returns and idiosyncratic volatility puzzle: Evidence from the chinese stock market. *Emerging Markets Finance and Trade*, *55*(5), 1184–1198.

Rahman, S. (2019). Discretionary tone, annual earnings and market returns: Evidence from uk interim management statements. *International Review of Financial Analysis*, *65*, 101384.

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 29–48).

Rehman, A., & Baloch, Q. B. (2016). Evaluating pakistan's mutual fund performance: Validating through capm and fama french 3-factor model. *Journal of Managerial Sciences*, *10*(1).

Rupande, L., Muguto, H. T., & Muzindutsi, P.-F. (2019). Investor sentiment and stock return volatility: Evidence from the johannesburg stock exchange. *Cogent Economics & Finance*, *7*(1), 1600233.

Santos, M. F., Cortez, P., Pereira, J., & Quintela, H. (2006). Corporate bankruptcy prediction using data mining techniques. *WIT transactions on information and communication technologies*, *37*.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425–442.

Shi, Y., Zhu, L., Li, W., Guo, K., & Zheng, Y. (2019). Survey on classic and latest textual sentiment analysis articles and techniques. *International Journal of Information Technology & Decision Making*, *18*(04), 1243–1287.

Shyklo, A. (2017). Simple explanation of zipf's mystery via new rank-share distribution, derived from combinatorics of the ranking process. *Derived from Combinatorics of the Ranking Process (February 15, 2017)*.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

Singhvi, S. S., et al. (1968). Corporate disclosure through annual reports in the united states of america and india. *Journal of Finance*, *23*(3), 551–552.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Suwanna, T. (2012). Impacts of dividend announcement on stock return. *Procedia-Social and Behavioral Sciences*, *40*, 721–725.

Sydserff, R., & Weetman, P. (2002). Developments in content analysis: a transitivity index and diction scores. *Accounting, Auditing & Accountability Journal*.

Tan, Z., Yan, Z., & Zhu, G. (2019). Stock selection with random forest: An exploitation of excess return in the chinese stock market. *Heliyon*, *5*(8), e02310.

Tavcar, L. R. (1998). Make the md&a more readable. *The CPA Journal*, *68*(1), 10.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), 1139–1168.

Welch, I. (2017). Stock return outliers. *Available at SSRN 3068347*.

Wisniewski, T. P., & Yekini, L. S. (2015). Stock market returns and the content of annual report narratives. In *Accounting forum* (Vol. 39, pp. 281–294).

Woodfield, T. J., & Irvine, C. (2006). Predicting workers' compensation insurance fraud using sas enterprise miner 5.1 and sas text miner. *Data Mining and Predictive Modeling. SAS Institute Inc., Irvine, CA*.

Wu, G. G.-R., Hou, T. C.-T., & Lin, J.-L. (2019). Can economic news predict taiwan stock market returns? *Asia Pacific management review*, *24*(1), 54–59.

Wujec, M. A. (2021). Sentiment analysis of current reports texts with use of cumulative abnormal return and deep neural network.

Yang, Q., Li, L., Zhu, Q., & Mizrach, B. (2017). Analysis of us sector of services with a new fama-french 5-factor model. *Applied Mathematics*, *8*(9), 1307–1319.

Yekini, L. S., Wisniewski, T. P., & Millo, Y. (2016). Market reaction to the positiveness of annual report narratives. *The British Accounting Review*, *48*(4), 415–430.

Ying, Q., Yousaf, T., Akhtar, Y., Rasheed, M. S., et al. (2019). Stock investment and excess returns: a critical review in the light of the efficient market hypothesis. *Journal of Risk and Financial Management*, *12*(2), 97.